

Visual Saliency Prediction Based on Deep Learning

By

© Bashir Ghariba, BSc, MSc

A thesis submitted to the School of Graduate Studies In partial
fulfillment of the requirements for the degree of

Doctor of Philosophy

Faculty of Engineering and Applied Science

Memorial University of Newfoundland

October 2020

St. John's, Newfoundland and Labrador

Dedication

I dedicate this thesis to

my grandmothers, my parents, my brothers, my sisters,

my wife, and my children for their constant support.

Abstract

The Human Visual System (HVS) has the ability to focus on specific parts of a scene, rather than the whole image. Human eye movement is also one of the primary functions used in our daily lives that helps us understand our surroundings. This phenomenon is one of the most active research topics in the computer vision and neuroscience fields. The outcomes that have been achieved by neural network methods in a variety of tasks have highlighted their ability to predict visual saliency. In particular, deep learning models have been used for visual saliency prediction. In this thesis, a deep learning method based on a transfer learning strategy is proposed (Chapter 2), wherein visual features in the convolutional layers are extracted from raw images to predict visual saliency (e.g., saliency map). Specifically, the proposed model uses the VGG-16 network (i.e., Pre-trained CNN model) for semantic segmentation. The proposed model is applied to several datasets, including TORONTO, MIT300, MIT1003, and DUT-OMRON, to illustrate its efficiency. The results of the proposed model are then quantitatively and qualitatively compared to classic and state-of-the-art deep learning models.

In Chapter 3, I specifically investigate the performance of five state-of-the-art deep neural networks (VGG-16, ResNet-50, Xception, InceptionResNet-v2, and MobileNet-v2) for the task of visual saliency prediction. Five deep learning models were trained over the SALICON dataset and used to predict visual saliency maps using four standard datasets, namely TORONTO, MIT300, MIT1003, and DUT-OMRON. The results indicate that the

ResNet-50 model outperforms the other four and provides a visual saliency map that is very close to human performance.

In Chapter 4, a novel deep learning model based on a Fully Convolutional Network (FCN) architecture is proposed. The proposed model is trained in an end-to-end style and designed to predict visual saliency. The model is based on the encoder-decoder structure and includes two types of modules. The first has three stages of inception modules to improve multi-scale derivation and enhance contextual information. The second module includes one stage of the residual module to provide a more accurate recovery of information and to simplify optimization. The entire proposed model is fully trained from scratch to extract distinguishing features and to use a data augmentation technique to create variations in the images. The proposed model is evaluated using several benchmark datasets, including MIT300, MIT1003, TORONTO, and DUT-OMRON. The quantitative and qualitative experiment analyses demonstrate that the proposed model achieves superior performance for predicting visual saliency.

In Chapter 5, I study the possibility of using deep learning techniques for Salient Object Detection (SOD) because this work is slightly related to the problem of Visual saliency prediction. Therefore, in this work, the capability of ten well-known pre-trained models for semantic segmentation, including FCNs, VGGs, ResNets, MobileNet-v2, Xception, and InceptionResNet-v2, are investigated. These models have been trained over an ImageNet dataset, fine-tuned on a MSRA-10K dataset, and evaluated using other public datasets, such as ECSSD, MSRA-B, DUTS, and THUR15k. The results illustrate the superiority of

ResNet50 and ResNet18, which have Mean Absolute Errors (MAE) of approximately 0.93 and 0.92, respectively, compared to other well-known FCN models.

Finally, conclusions are drawn, and possible future works are discussed in chapter 6.

Acknowledgements

First, I am thankful to Almighty ALLAH; without His blessings it would have been impossible to finish this thesis. I would also like to express my gratitude and thanks to the the Higher Education Ministry of the Libyan government and the Elmergib University, Khoms, Libya for providing me the scholarship to continue my PhD studies. I would also like to give special thanks to all the teachers and professors who have taught me during my various study stages.

A special thanks is given to my family back in Libya, especially my grandmother and father who did not live long enough to see this time, but their memories will always be with me. Also deep thanks to my dear mother, brothers, sisters, relatives, and friends. I would like to express my appreciation to my wife (Eman) and my children (Muftah, Mohamed, Monea, Mera, and little baby: Maysem) for their continued support and encouragement during my PhD journey.

I would like to express my sincere gratitude and deepest appreciation to my research supervisor, Dr. Mohamed Shehata and my Co-supervisor, Dr. Peter McGuire for their broad knowledge, support, kind attitude, research guidance, suggestions, and inspiration throughout the period of this work that made possible the successful completion of this

thesis. Also special thanks to my supervisor committee member, Dr. Theodore S. Norvell for his cooperation.

Co-authorship Statement

I, Bashir Ghariba, have the principal author status for all manuscripts included in this thesis. The other authors, Dr. Mohamed Shehata and Dr. Peter McGuire, made valuable contributions that promoted the development of this work. The list of the peer-reviewed manuscripts that included in this thesis are described as follows:

- 1- Bashir Ghariba, Mohamed S. Shehata, and Peter McGuire, "Visual Saliency Prediction Based on Deep Learning," *Information*, vol. 10, no. 8, pp. 257-271, August 2019.
- 2- Bashir Ghariba, Mohamed S. Shehata, and Peter McGuire, "A Novel Fully Convolutional Network for Visual Saliency Prediction," *PeerJ Computer Science* 6 (July 2020): e280.
- 3- Bashir Ghariba, Mohamed S. Shehata, and Peter McGuire, "Salient Object Detection Using Semantic Segmentation Techniques," *International Journal of Computational Vision and Robotics*, September 2020.
- 4- Bashir Ghariba, Mohamed S. Shehata, and Peter McGuire, "Performance Evaluation of Pre-Trained CNN Models for Visual Saliency Prediction," in *Proc. IEEE Canadian Conference on Electrical and Computer Engineering (CCECE 2020)*, PP. 1- 4, August 30 – September 02, 2020.

Table of Contents

Dedication	II
Abstract	III
Acknowledgements	V
Co-authorship Statement.....	VI
Table of Contents	VII
List of Tables	XIII
List of Figures	XV
List of Acronyms	XVI
Chapter 1. Introduction	1
1.1 Research Motivation.....	1
1.2 Research Objectives	1
1.3 Thesis Organization.....	2
1.4 Contributions	3
1.5 Literature Review	5
1.5.1 Background.....	5

1.5.2 Bottom-Up Strategy.....	6
1.5.3 Top-Down Strategy	6
1.5.4 Conventional Saliency Models (Non-Deep Learning Era)	7
1.5.5 Deep Learning Saliency Models	8
References	12
Chapter 2. Visual Saliency Prediction Based on Deep Learning.....	18
Abstract	18
2.1 Introduction	18
2.2 The Proposed Method	22
2.2.1 The VGG-16 Network Architecture	22
2.2.2 Visual Saliency Prediction Model	25
2.3 Materials and Methods	28
2.3.1 Model Training	28
2.3.2 Model Testing.....	30
2.3.3 Datasets.....	31
2.3.4 Evaluation Metrics.....	33
2.4 Experimental Results.....	36
2.4.1 Quantitative Comparison of the Proposed Model with Other State-of-the-Art Models	36
2.4.2 Qualitative Comparison of the Proposed Model with Other State-of-the-Art Models	39

2.5 Conclusion.....	41
References	42
Chapter 3. Performance Evaluation of Pre-Trained CNN Models for Visual Saliency	
Prediction	48
Abstract	48
3.1 Introduction	48
3.2 Materials and Methods	50
3.2.1 Semantic Segmentation	50
3.2.2 Pre-Trained Models	50
3.2.3 Pre-Trained Models Specification	52
3.2.4 Datasets.....	52
3.2.5 Evaluation Metrics.....	54
3.3 Experimental Results.....	55
3.3.1 Models Training	55
3.3.2 Visual Results.....	56
3.3.3 Numerical Results	58
3.4 Conclusion.....	60
References	61
Chapter 4. A Novel Fully Convolutional Network for Visual Saliency Prediction.....	
Abstract	64

4.1 Introduction	65
4.2 Related work	69
4.3 Material and Methods.....	71
4.3.1 Proposed Model.....	71
4.3.2 Semantic Segmentation	75
4.3.3 Datasets.....	75
4.3.4 Evaluation Metrics.....	77
4.4 Experimental Results.....	80
4.4.1 Model Training.....	80
4.4.2 Model Testing.....	82
4.5 Discussion	83
4.5.1 Quantitative Comparison of the Proposed Model with Other Advanced Models	83
4.5.2. Qualitative Comparison of the Proposed Model with Other Advanced Models	87
4.5.3 Ablation Study.....	88
4.6 Conclusions	90
References	92
Chapter 5. Salient Object Detection Using Semantic Segmentation Techniques.....	100
Abstract	100

6.2 Future Work	138
References	140

List of Tables

1.1	A timeline of visual saliency based on deep-learning since 2014	09
2.1	Comparison of the quantitative scores of several models on TORONTO dataset.....	37
2.2	Comparison of the quantitative scores of several models on MIT300 dataset.....	38
2.3	Comparison of the quantitative scores of several models on MIT1003 dataset.....	38
2.4	Comparison of the quantitative scores of several models on DUT-OMRON dataset.....	38
2.5	Model predication results (i.e., Global accuracy) on several datasets (TORONTO, MIT300, MIT1003, and DUT-OMRON)	39
3.1	Pre-Trained Model Parameters.	52
3.2	The most important parameters through the training stage of all pre-trained models.....	56
3.3	SIM and NSS metrics obtained from the state-of-the-art pre-trained models for the A : TORONTO, B : MIT300, C : MIT1003, and D : DUT-OMRON datasets.....	59
3.4	AUC-Judd and AUC-Borji metrics obtained from the state-of- the-art pre-trained models for A : TORONTO, B : MIT300, C : MIT1003, and D : DUT-OMRON datasets ...	59
4.1	Configuration of the proposed model.....	73
4.2	Configuration of inception and residual modules.....	74
4.3	Comparison of the quantitative scores of several models on TORONTO dataset.....	84
4.4	Comparison of the quantitative scores of several models on MIT300 dataset.....	85
4.5	Comparison of the quantitative scores of several models on MIT1003 dataset.....	85
4.6	Comparison of the quantitative scores of several models on DUT-OMRON dataset.....	86
4.7	Properties of the proposed model and ten visual saliency models.....	86
4.8	Different FCN models applied in this study.....	89

5.1	ResNet-18 Architecture.....	110
5.2	The most important parameters through the training stage of all pre-trained models.....	119
5.3	Several parameters that influence the training and testing of different models in this study.....	120
5.4	Comparison of the quantitative scores of different models on ECSSD and MSRA-B....	122
5.5	Comparison of quantitative scores of different models on DUTS and THUR15k.....	123
5.6	Evolution metrics of selected trained models for SOD against noise in the selected datasets.....	127

List of Figures

2.1	General Structure of VGG-16:(a) Convolution layers of VGG-16, and (b) Data flow in VGG-16....	23
2.2	Architecture of the Proposed model for visual saliency prediction.....	26
2.3	Value of validation accuracy and loss as a function of epochs.....	29
2.4	Model testing :(a) TORONTO and MIT300 datasets, and (b) MIT1003 and DUT- OMRON dataset.....	31
2.5	The saliency maps obtained from the proposed model and five other state-of-the-art models for a sample image from the TORONTO, MIT300, MIT1003, and DUT-OMRON datasets.....	40
3.1	Visual saliency maps predicted by state-of-the-art pre-trained models for images drawn from the TORONTO, MIT300, MIT1003, and DUT-OMRON datasets.....	57
4.1	Architecture of the proposed model.....	73
4.2	Architecture of (a) Inception and (b) residual modules.....	74
4.3	Value of validation accuracy (a) and loss as a function of epochs (b).....	82
4.4	Saliency maps obtained from the proposed model and five advanced models for a sample image from the TORONTO, MIT300, MIT1003, and DUT-OMRON datasets.....	88
5.1	General Structure of VGG-16 network: (a) Convolution layers, and (b) data flow.....	107
5.2	Data flow in VGG-19 Network.....	109
5.3	ResNet-50 Network architecture.....	110
5.4	Main block of MobileNet-v2 Network.....	111
5.5	The architecture of Xception Network.....	112
5.6	The architecture of InceptionResnet-v2 Network.....	113
5.7	FCN models architecture.....	114
5.8	Comparison of salient object detection models on: (a) ECSSD and MSRA-B; (b) DUTS and THUR15k datasets.....	125

List of Acronyms

ADAM	Adaptive Moment Estimation
AUC	Area Under ROC Curve
AWS	Adaptive Whitening Saliency
CNNs	Convolutional Neural Networks
Conv	Convolutional Operation
CVPR	Computer Vision and Pattern Recognition
Decon	Deconvolution Operation
DL	Deep learning
Fbw	Weighted F_B measure
fc	fully connected layer
FCNs	Fully Convolutional Networks
FP	False Positive
GBVS	Graph-Based Visual Saliency
GT	Ground Truth
HVA	Human Visual Attention
HVS	Human Visual System
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IoU	Intersection over Union
JPG	Joint photographic Experts Group
MAE	Mean Absolute Error

MIT	Massachusetts Institute of Technology
MS COCO	Microsoft Common Objects in Context
NSS	Normalized Scanpath Saliency
PNG	Portable Network Graphics
PR	Precision-Recall
ReLU	Rectified Linear Unit
RMSProp	Root Mean Square propagation
ROC	Receiver Operating Characteristics
SALICON	Saliency in Context
SGDM	Stochastic Gradient Descent with Momentum
SIM	Similarity Metrics
SOD	Salient Object Detection
SR	Spectral Residual Saliency
TP	True Positive
VGG	Visual Geometry Group

Chapter 1. Introduction

1.1 Research Motivation

Human Visual Attention (HVA) is a very important function in the human visual system. To focus on a part of the scene instead to focus on the whole scene, we need to move our eyes to that specific place. The phenomenon of visual attention has been studied for over a century [1]. Recent experiments suggest that attention is required for us to perceive anything at all [1]. During the day, we scan the visual spotlight surrounding the environment, targeting things like words, faces, images on a package, reading books, and a variety of other objects. Recently, neural networks have been used to deal with visual attention phenomena (i.e., visual saliency prediction) [2]. Despite these tremendous efforts and immense progress, there continues to be room for improvement in terms of datasets, evaluation measures, cognitive study, analysis of deep network models, applications, and the prediction accuracy of various models [2].

1.2 Research Objectives

Despite significant improvements to the problem of visual saliency prediction based on deep learning techniques, recent models still cannot fully understand high-level semantics (i.e., the semantic gap). Therefore, the main objective of this research work is to decrease the semantic gap for visual saliency prediction, thus reaching human visual system performance.

In order to achieve the main objective of this research, the following specific objectives are defined:

- 1- Investigate the use of semantic segmentation techniques based on the encoder-decoder structure. This structure employed the pre-trained network (i.e., VGG-16) for predicting human eye fixation.
- 2- Study the performance of five pre-trained CNN models to solve the problem of visual saliency prediction.
- 3- Develop a novel model based on the encoder-decoder structure. This model includes an addition inception module and skip connections with a residual module. The former module will improve multi-scale inference and enrich contextual information, while the latter contributes to the recovery of more detailed information, simplifies optimization, and avoids the vanishing gradient problem.
- 4- Investigate ten pre-trained CNN models for salient Object Detection (SOD). These models including FCNs(8s,16s,32s), VGGs (16,19), ResNets (18,50), MobileNet-v2, Xception, and InceptionResNet-v2.

1.3 Thesis Organization

This thesis is a paper-based one that contains six chapters. Chapter 1 is the introduction, which includes a description of the research motivation, research objectives, thesis structure, contributions, and a literature review. In Chapter 2, a deep learning model based on a pre-trained network (VGG-16) for visual saliency prediction is proposed. In Chapter 3, the performance of several CNN pre-trained models for visual saliency prediction is

investigated. Chapter 4 describes a novel model based on a Fully Convolutional Network (FCN) (full-training), to predict visual saliency. In Chapter 5, 10 pre-trained CNN models based on semantic segmentation techniques for Salient Object Detection (SOD) are then investigated. Finally, the overall conclusions are drawn and future work is outlined in Chapter 6.

1.4 Contributions

This thesis presents the following contributions:

1. A deep learning model based on the semantic segmentation technique for visual saliency prediction is proposed. The VGG-16 network is a pre-trained network, which is appropriate for achieving tasks that do not have enough datasets for model training. The proposed model was trained on a well-known dataset (SALICON) and was also evaluated on other datasets, including TORONTO, MIT300, MIT1003, and DUT-OMRON. The trained model is able to predict visual saliency (e.g., saliency map) with reasonable accuracy with respect to other state-of-the-art models.
2. The performance of five state-of-the-art deep neural networks (VGG-16, ResNet-50, Xception, InceptionResNet-v2, and MobileNet-v2) for the task of visual saliency prediction are investigated. In this work, five deep learning models are trained over the SALICON dataset and then used to predict visual saliency maps using four standard datasets, namely TORONTO, MIT300, MIT1003, and DUT-OMRON. The results indicate that the ResNet-50 model outperforms the other four and provides a visual saliency map that is very close to human performance.

3. A novel model based on a Fully Convolutional Network (FCN) architecture is proposed. The proposed model is trained in an end-to-end style and designed to predict visual saliency. The model is based on the encoder-decoder structure and includes two types of modules. The first has three stages of inception modules to improve multi-scale derivation and enhance contextual information. The second module includes one stage of the residual module to provide a more accurate recovery of information and simplify optimization. The entire proposed model is fully trained from scratch to extract distinguishing features and to use a data augmentation technique to create variations in the images. The proposed model is evaluated using several benchmark datasets, including: MIT300, MIT1003, TORONTO, and DUT-OMRON.

4. Finally, the work described above is extended for Salient Object Detection (SOD). While this task differs from visual saliency prediction, it is related. In this work, the capability of several well-known pre-trained models for semantic segmentation, including FCNs, VGGs, ResNets, MobileNet-v2, Xception, and InceptionResNet-v2, are investigated. These models have been trained over an ImageNet dataset, fine-tuned on a MSRA-10K dataset, and evaluated using other public datasets, including ECSSD, MSRA-B, DUTS, and THUR15k. The results illustrate the superiority of ResNet50 and ResNet18, compared to other well-known FCN models. Moreover, the most robust model against noise is ResNet50, whereas VGG-16 is the most sensitive, relative to other state-of-the-art models.

1.5 Literature Review

1.5.1 Background

In this section, a brief review of related work on visual saliency prediction is provided. Then, the architecture used in the task of visual saliency prediction is summarized. There are two types of architecture of visual saliency prediction models: the first is based on classical methods (non-deep learning), whereas the second is based on deep learning methods. In the last few years, several models have been proposed for the prediction of visual saliency. The pattern that predicted using visual saliency prediction can be defined as a saliency map. The saliency map illustrates that the location of human attention is a unique area within the whole image [3]. Importantly, the purpose of a saliency map is to change the representation of an image to a smooth image that is more meaningful and easier to analyze [4].

Generally, Human Visual Attention (HVA) is based on two strategies; bottom-up and top-down visual attention. Bottom-up models mainly employ low-level cues, such as color, intensity, and texture. Additionally, bottom-up strategies try to select regions that show prominent characteristics from their surroundings. In contrast, top-down approaches are task-oriented and try to locate a target object from a specific category and are, therefore, dependent on the features of the interesting object [5, 6]. More details about the two strategies will be explained in the next sections.

1.5.2 Bottom-Up Strategy

The bottom-up visual saliency strategy is an important module of human visual attention and is based on the location of visually important features in the image. This phenomenon of visual attention is inspired by a biological vision, following the feature integration of Treisman's theory [7]. This theory suggests that when observing a stimulus, features are recorded early, automatically, and in parallel across the visual field, while objects are specified separately and only at a later stage, which requires localized attention. Moreover, this module decomposes visual input into separate low-level feature maps, such as color, contrast, and orientation. For every single feature, a different map is calculated and normalized. Therefore, a saliency map is formed by the weighted collection of all maps. Highs in the map reflect the attention (i.e., saliency) [8]. Over the past several years, the results of bottom-up modules have been evaluated using mathematical and statistical tools. These models also have been evaluated using eye movement data provided by the gaze location of viewers [9].

1.5.3 Top-Down Strategy

In contrast, the top-down module is driven by a task. This module uses prior knowledge, rewards, or expectations as high-level visual factors to recognize the target of interest. Several top-down saliency models have been proposed [7]. Top-down modules are primarily investigated by cueing experiments, in which a hint brings one's attention to the target. This hint could be what the target is and where it is located. Oliva et al. presented a top-down visual search model using a Bayesian framework [8]. Gao et al. also proposed a

top-down model based on decision-theoretic models [3]. They explained a top-down saliency as a classification task wherein locations where a target could be differentiated from non-targets with the least amount of error are classified as salient.

1.5.4 Conventional Saliency Models (Non-Deep Learning Era)

Several classic non-deep learning saliency models have been proposed, all of which are based on bottom-up strategies. These models were introduced by Treisman and Gelade in 1980 [9], the computational architecture was developed by Koch and Ullman in 1985 [10], and the bottom-up model of Itti et al. was proposed in 1998 [11]. These models are based on feature map extraction, such as color, intensity, texture, and orientation. Itti et al. proposed a model which was able to predict human behavior in visual search functions, explain robustness to image noise, traffic sign detection, and predict where humans look while viewing of images and videos [12, 13].

Many models have been introduced to predict static saliency, such as Attention for Information Maximization (AIM) [14], Graph-based Visual Saliency (GBVS) [15], Spectral Residual saliency (SR) [16], Boolean Map based Saliency (BMS) [17], Saliency Using Natural statistics (SUN) [18], Adaptive Whitening Saliency (AWS) [19], and the Judd et al. model [5]. Many models have also been proposed for extracting dynamic saliency (i.e., video images), such as AWS-D [20], Xu et al. [21], Rudoy et al. [22], OBDL [23], and PQFT [24]. Most of these static and dynamic models are inspired by biological vision.

1.5.5 Deep Learning Saliency Models

Deep learning (DL) has become the fastest-growing trend in big data analysis and is widely considered as one of the top ten breakthrough technologies of the year 2013 [25]. The success of deep learning has brought with it a new wave of saliency models that perform much better than classical saliency systems based on hand-crafted features. Researchers now utilize available deep architectures that are properly trained to recognize the scene and then reuse them to predict visual saliency. In general, deep learning techniques require a large dataset to deliver high performance. However, these large-scale datasets for fixation are not enough. Therefore, deep saliency models are pre-trained on a huge dataset and are then trained (i.e. fine-tuned) on smaller scale mouse click or eye movement datasets. Subsequently, this method allows for models to reuse the knowledge (i.e., Semantic knowledge) that is learned in CNNs and transfer it to the visual saliency task.

1.5.5.1 Static Saliency Models

In general, static saliency models have been in use since 2014 and form the basis of modern models. Table 1.1 explains the timeline for the static visual saliency models based on deep learning techniques.

Table 1.1: A timeline of static visual saliency based on deep learning since 2014.

2014	2015	2016	2017	2018	2019
-eDN [26] -Deep Gaze I [27]	-Mr-CNN [28] -DeepFix [29] -SALICON [30]	-JuntingNet [31] -SalNet [32] -PDP [33] -DSCLRCN [34] -ML-Net [35]	-SalGAN [36] -Deep Gaze II [37] -iSEEL [38]	-EML-NET [39] -SAM ResNet [40] -DVA [41]	-MSI-Net [42] -SDS [43]

1.5.5.2 Dynamic Saliency Models

In general, dynamic saliency models have high computational and memory requirements. Therefore, working with models of video saliency prediction is a more challenging task than image saliency prediction. Regardless, there has been an increasing interest in video saliency over the past few years driven by its applications (e.g., image, video summarization, and video captioning).

Conventionally, video saliency models pair bottom-up feature extraction with an ad-hoc motion estimation that can be achieved either by means of optical flow or feature tracking. Moreover, deep video saliency models learn the whole process end-to-end. In these works, the dynamic characteristics are modeled in one of two ways: 1) adding temporal information to the CNN, or 2) developing a dynamic structure using LSTMs. In addition, there are many models that have been developed for visual saliency prediction in videos, such as:

1. **Two-stream network:** As one of the first attempts, Bak et al. [44] applied a two stream (five years each) CNN architecture for video saliency prediction. RGB frames and motion maps were fed into the two streams.
2. **Chaabouni et al.:** These authors employed transfer learning to adapt a previously trained deep network for saliency prediction in natural videos. They trained a five layered CNN on RGB color planes and residual motion for each video frame. However, their models use only the very short-term temporal relations of two consecutive frames [45].
3. **Bazzani et al.:** A recurrent mixture density network is proposed for saliency prediction [46]. The input clip of 16 frames is inserted to a 3D CNN, the output of which becomes the input for a LSTM. Finally, a linear layer projects the LSTM representation to a Gaussian mixture model, which describes the saliency map.
4. **OM-CNN:** Proposed by Jiang et al., the Object-to-Motion CNN model includes two subnets of objectness and motion that are trained end-to-end. Objectness and object motion information are used to predict intraframe saliency of videos [47].
5. **Leifman et al.:** The authors introduced a novel Depth-Aware Video Saliency approach to predict human focus of attention when viewing RGBD videos on regular 2D screens [48].
6. **Gorji & Clark:** These authors proposed a multi-stream Convolutional Long Short-Term Memory network (ConvLSTM) structure which augments state-of-the-art in static saliency models with dynamic Attentional Push. Their network contains a saliency pathway and three push pathways [49].

7. **ACLNet:** The Attentive CNN-LSTM Network augments a CNN-LSTM with a supervised attention mechanism to enable fast end-to-end saliency learning. The attention mechanism encodes static saliency information, allowing LSTM to focus on learning a more flexible temporal saliency representation across successive frames [50].
8. **SG-FCN** Sun et al. proposed a robust deep model that utilizes memory and motion information to capture salient points across successive frames. The memory information enhanced the model's generalization because changes between two adjacent frames are limited within a certain range, and hence the corresponding fixations should remain correlated [51].

References

1. Allport, A., Visual attention. 1989.
2. Wang, W. and J. Shen, Deep visual attention prediction. *IEEE Transactions on Image Processing*, 2018. **27**(5): p. 2368-2378.
3. Gao, D. and N. Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. in *Advances in neural information processing systems*. 2005.
4. Cornia, M., et al., Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2018. **14**(2): p. 48.
5. Judd, T., et al. Learning to predict where humans look. in *2009 IEEE 12th international conference on computer vision*. 2009. IEEE.
6. Rensink, R.A., The dynamic representation of scenes. *Visual cognition*, 2000. **7**(1-3): p. 17-42.
7. Borji, A. and L. Itti, State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 2012. **35**(1): p. 185-207.
8. Oliva, A., et al. Top-down control of visual attention in object detection. in *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*. 2003. IEEE.
9. Treisman, A.M. and G. Gelade, A feature-integration theory of attention. *Cognitive psychology*, 1980. **12**(1): p. 97-136.
10. Koch, C. and S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, in *Matters of intelligence*. 1987, Springer. p. 115-141.

11. Itti, L., C. Koch, and E. Niebur, A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1998(11): p. 1254-1259.
12. Parkhurst, D., K. Law, and E. Niebur, Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 2002. **42**(1): p. 107-123.
13. Peters, R.J., et al., Components of bottom-up gaze allocation in natural images. *Vision research*, 2005. **45**(18): p. 2397-2416.
14. Bruce, N. and J. Tsotsos. Saliency based on information maximization. in *Advances in neural information processing systems*. 2006.
15. Harel, J., C. Koch, and P. Perona. Graph-based visual saliency. in *Advances in neural information processing systems*. 2007.
16. Hou, X. and L. Zhang. Saliency detection: A spectral residual approach. in *2007 IEEE Conference on computer vision and pattern recognition*. 2007. IEEE.
17. Zhang, J. and S. Sclaroff. Saliency detection: A boolean map approach. in *Proceedings of the IEEE international conference on computer vision*. 2013.
18. Zhang, L., et al., SUN: A Bayesian framework for saliency using natural statistics. *Journal of vision*, 2008. **8**(7): p. 32-32.
19. Garcia-Diaz, A., et al., Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 2012. **30**(1): p. 51-64.
20. Leboran, V., et al., Dynamic whitening saliency. *IEEE transactions on pattern analysis and machine intelligence*, 2017. **39**(5): p. 893-907.

21. Xu, M., et al., Learning to detect video saliency with HEVC features. *IEEE Transactions on Image Processing*, 2017. **26**(1): p. 369-385.
22. Rudoy, D., et al. Learning video saliency from human gaze using candidate selection. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013.
23. Hossein Khatoonabadi, S., et al. How many bits does it take for a stimulus to be salient? in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
24. Guo, C., Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 2008. IEEE.
25. LeCun, Y., et al., Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. **86**(11): p. 2278-2324.
26. Vig, E., M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.
27. Kümmerer, M., L. Theis, and M. Bethge, Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014.
28. Liu, N., et al. Predicting eye fixations using convolutional neural networks. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

29. Kruthiventi, S.S., K. Ayush, and R.V. Babu, DeepFix: A fully convolutional neural network for predicting human eye fixations. CoRR abs/1510.02927 (2015). arXiv preprint arXiv:1510.02927, 2015.
30. Huang, X., et al. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. in Proceedings of the IEEE International Conference on Computer Vision. 2015.
31. Pan, J. and X. Giró-i-Nieto, End-to-end convolutional network for saliency prediction. arXiv preprint arXiv:1507.01422, 2015.
32. Pan, J., et al. Shallow and deep convolutional networks for saliency prediction. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
33. Jetley, S., N. Murray, and E. Vig. End-to-end saliency mapping via probability distribution prediction. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
34. Liu, N. and J. Han, A deep spatial contextual long-term recurrent convolutional network for saliency detection. IEEE Transactions on Image Processing, 2018. **27**(7): p. 3264-3274.
35. Cornia, M., et al. A deep multi-level network for saliency prediction. in 2016 23rd International Conference on Pattern Recognition (ICPR). 2016. IEEE.
36. Pan, J., et al., Salgan: Visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081, 2017.

- 37. Kummerer, M., et al. Understanding low-and high-level contributions to fixation prediction. in Proceedings of the IEEE International Conference on Computer Vision. 2017.
- 38. Tavakoli, H.R., et al., Exploiting inter-image similarity and ensemble of extreme learners for fixation prediction using deep features. *Neurocomputing*, 2017. **244**: p. 10-18.
- 39. Jia, S. and N.D. Bruce, Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, 2020: p. 103887.
- 40. Cornia, M., et al., Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 2018. **27**(10): p. 5142-5154.
- 41. Wang, W. and J. Shen, Deep visual attention prediction. *IEEE Transactions on Image Processing*, 2017. **27**(5): p. 2368-2378.
- 42. Kroner, A., et al., Contextual encoder-decoder network for visual saliency prediction. *Neural Networks*, 2020.
- 43. Li, Y. and X. Mou, Saliency detection based on structural dissimilarity induced by image quality assessment model. *Journal of Electronic Imaging*, 2019. **28**(2): p. 023025.
- 44. Bak, C., et al., Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE Transactions on Multimedia*, 2018. **20**(7): p. 1688-1698.

45. Chaabouni, S., J. Benois-Pineau, and C.B. Amar. Transfer learning with deep networks for saliency prediction in natural video. in 2016 IEEE International Conference on Image Processing (ICIP). 2016. IEEE.
46. Bazzani, L., H. Larochelle, and L. Torresani, Recurrent mixture density network for spatiotemporal visual attention. arXiv preprint arXiv:1603.08199, 2016.
47. Jiang, L., M. Xu, and Z. Wang, Predicting video saliency with object-to-motion CNN and two-layer convolutional LSTM. arXiv preprint arXiv:1709.06316, 2017.
48. Leifman, G., et al. Learning gaze transitions from depth to improve video saliency estimation. in Proceedings of the IEEE International Conference on Computer Vision. 2017.
49. Gorji, S. and J.J. Clark. Going from image to video saliency: Augmenting image salience with dynamic attentional push. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
50. Wang, W., et al. Revisiting video saliency: A large-scale benchmark and a new model. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
51. Sun, M., et al., Sg-fcn: A motion and memory-based deep learning model for video saliency detection. IEEE transactions on cybernetics, 2018(99): p. 1-12.

Chapter 2. Visual Saliency Prediction Based on Deep Learning

Abstract

Human eye movement is one of the most important functions for understanding our surroundings. When a human eye processes a scene, it quickly focuses on dominant parts of the scene, commonly known as visual saliency detection or visual attention prediction. Recently, neural networks have been used to predict visual saliency. This work proposes a deep learning encoder-decoder architecture, based on a transfer learning technique, to predict visual saliency. In the proposed model, visual features are extracted through convolutional layers from raw images to predict visual saliency. In addition, the proposed model uses the VGG-16 network for semantic segmentation, which uses a pixel classification layer to predict the categorical label for every pixel in an input image. The proposed model is applied to several datasets, including TORONTO, MIT300, MIT1003, and DUT-OMRON, to illustrate its efficiency. The results of the proposed model are quantitatively and qualitatively compared to classic and state-of-the-art deep learning models. Using the proposed deep learning model, a global accuracy of up to 96.22% is achieved for the prediction of visual saliency.

2.1 Introduction

Humans have a strong ability to pay attention to a specific part of an image instead of processing the entire image. This phenomenon of visual attention has been studied for over a century [1]. Visual attention is defined as the processes that enable an observer to focus

on selected aspects of the retinal image over non-selected aspects. In other words, visual attention refers to a set of cognitive procedures that select relevant information and filter out irrelevant information from cluttered visual scenes. The task of visual attention prediction is a popular research area in the computer vision and neuroscience fields. In general, Human Visual Attention (HVA) is based on two strategies: bottom-up and top-down visual attention. Bottom-up models mainly employ low-level cues, such as color, intensity, and texture. Additionally, the bottom-up strategy tries to select regions which show the prominent characteristics of their surroundings [2,3]. In contrast, top-down approaches are task-oriented and try to locate a target object from a specific category. They also depend on the features of the object of interest [4,5]. Accordingly, bottom-up and top-down approaches are mainly driven by the visual characteristics of a scene and the task of interest, respectively [6,7].

In the last few years, several models have been proposed for the prediction of human visual saliency, with the most common technique being a saliency map. Saliency maps illustrate that the location of human attention is focused on a particular area within the whole image [8–10]. In addition, a saliency map is an image that shows each pixel's unique quality. Importantly, the purpose of a saliency map is to change the representation of an image to a smooth image that is more meaningful and easier to analyze [11,12].

Deep Convolutional Neural Networks (CNNs) have been commonly used in the field of visual attention. This is because CNNs are strong visual models and they are able to learn features from a raw image dataset (low-level feature) and create a feature map (high-level

feature) [13,14]. This scenario describes how the human visual system can detect the location of visual attention. In the last few years, several deep learning models have been used to predict visual saliency points, most of which have achieved impressive performances compared to conventional methods [15–18]. The task of extracting a saliency map has further opened the door for several applications, especially in computer vision, including object detection, object recognition, scene classification, video understanding, and image compression [19].

This study aims to propose the application of a semantic segmentation model based on the VGG-16 network (see Section 2.2.1 for more details on the VGG-16 network) to predict human visual attention in the field of view. Specifically, the main objective of this research is to improve the accuracy of visual saliency prediction by proposing a fully convolutional neural network-based model. The proposed method that we used falls under the bottom-up category. Therefore, in the results section, we only compare our proposed method with relevant bottom-up methods (see Section 2.4.1 for more details on relevant methods).

The proposed model was developed based on the encoder-decoder architecture, wherein the fine-tuning strategy was applied in the encoder stage (i.e., VGG-16 model) [20]. More specifically, this study uses a VGG-16 model that was trained on more than a million images from the ImageNet database [20,21]. In addition, we trained the proposed model using SALICON images (see Section 2.3.3.1 for more details on the SALICON dataset) and their ground truth data [22], and evaluated the results over several datasets, including TORONTO, MIT300, MIT1003, and DUT-OMRON [23–25].

The contributions of this work can be summarized in the following points:

- (1) A deep learning architecture based on the VGG-16 network that is able to predict visual saliency is proposed. As opposed to the current state-of-the-art technique that uses three stages in the encoder/decoder architecture [26], the proposed network uses five encoder and decoder stages to produce a useful saliency map (e.g., visual saliency). This makes the proposed architecture more powerful for extracting more specific deep features.
- (2) The proposed model is the first to use a semantic segmentation technique within the encoder-decoder architecture to classify all image pixels into the appropriate class (foreground or background), where the foreground is most likely a salient object.
- (3) The proposed model is evaluated using four well-known datasets, including TORONTO, MIT300, MIT1003, and DUT-OMRON. The proposed model achieves a reasonable result, with a global accuracy of 96.22%.

To this end, the proposed method, based on the VGG-16 network, is described in Section 2.2; the materials and methods of the proposed model in Section 2.3; and the quantitative and qualitative experimental results obtained from the four datasets are explained in Section 2.4. Finally, we summarize our results in the conclusion and report potential future uses, applications, and improvements to this research in Section 2.5.

2.2 The Proposed Method

The proposed model is based on a semantic segmentation technique using the VGG-16 network. Hereby, we thoroughly explain all the important information about the VGG-16 network in the next sub-sections.

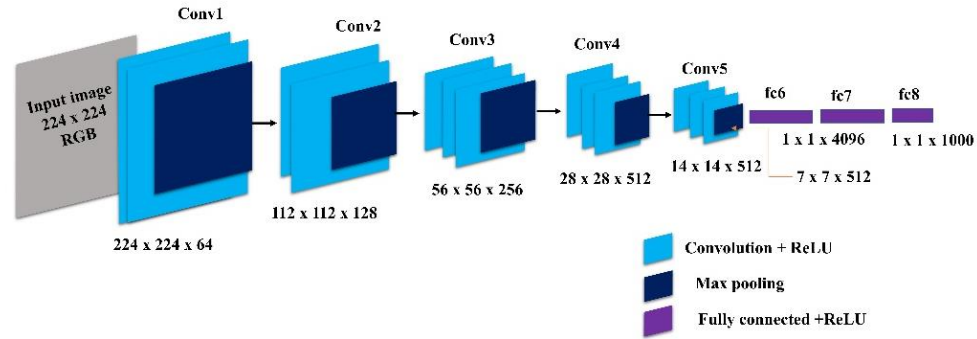
2.2.1 The VGG-16 Network Architecture

In this section, we describe the architecture of the proposed model. Our model architecture consists of encoder-decoder stages; the encoder stage has five convolutional blocks (conv1, conv2, conv3, conv4, and conv5). The encoder blocks are learned by down-sampling, which applies different receptive field sizes to create the feature maps. The decoder stage has also five deconvolution blocks (decon1, decon2, decon3, decon4, and decon5). The decoder blocks up-sample the feature maps and this creates an output the same size as the input image. The encoder blocks are adopted from a pretrained network called VGG-16 [14].

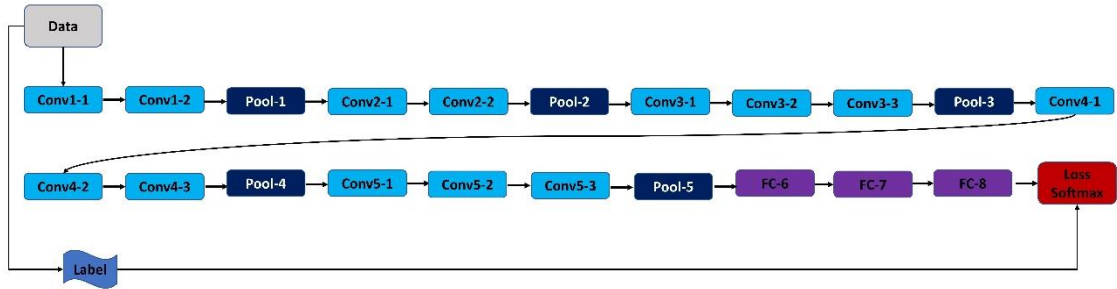
VGG-16 network was developed by Simonyan and Zisserman in the 2014 ILSVRC competition [13]. Generally, the VGG-16 network contains thirteen convolution layers, five pooling layers, and three fully connected layers [20]. The VGG-16 network is trained on more than a million images from the ImageNet database [21] and can classify images into 1000 object classes. The VGG-16 network has an image input size of 224×224 . Figure 2.1 (a, b) shows the general structure and the data flow through the VGG-16 network.

The major difference between VGG-16 network and the previous networks is the use of a series of convolution layers with small receptive fields (3×3) in the first layers instead of a

few layers. This results in fewer parameters and more nonlinearities in between, making the decision function more selective and the model easier for training [13].



(a)



(b)

Figure 2.1: General Structure of VGG-16 network: (a) Convolution layers of VGG-16 network, and (b) Data flow in VGG-16 network [13].

The input image is passed over a series of convolution layers with 3×3 convolutional filters. This is beneficial because the filter will capture the notation of the center, left/right, and up/down. The convolution stride is set to 1 pixel, whereas the padding is set to 1-pixel. Five max-pooling layers are used after convolution layers for the down-sampling operation (i.e., dimensionality reduction). Each max-pooling is also performed over 2×2 pixels, with

stride 2. In addition, three fully connected (fc) layers follow a series of convolution layers. In specific, the first two have 4096 channels each, and the third has 1000 channels. The structure of the fully connected layers is the same in all networks. The final layer is a soft-max layer that must have the same number of nodes as the output layer. The function of the soft-max layer is to map the non-normalized output to probability distribution through predicted output classes [13].

The convolutional neural network can be considered as the composition of several functions as:

$$f(x) = f_L (... f_2 (f_1 (x; w_1); w_2) ...), w_L), \quad (2.1)$$

where each function f_L takes as input a datum x_L and a parameter vector w_L and produces as output a datum x_{L+1} . The parameters $w = (w_1, ..., w_L)$ are learned from the input data for solving a specific problem, for example image classification. Moreover, there is a function called non-linear activation (i.e., not linear function), which is associated with the convolution layers. This function is also used to keep all the input value of the network as positive value. Equation (2.2) explains this concept.

$$y_{ijk} = \max(0, x_{ijk}) \quad (2.2)$$

There is another important operator also associated with the architecture of the VGG-16 network that is called the pooling operator. The purpose of this operator is reducing the dimension of the input volume (i.e., sub-sampling method) and preserving discriminant

information. There are several types of the operator, such as max-pooling, average-pooling, and sum-pooling. For instance, the output of a $p \times p$ max-pooling operator is:

$$y_{ijk} = \max\{y_{i'j'k} : i \leq i' < i + p, j \leq j' < j + p\}. \quad (2.3)$$

2.2.2 Visual Saliency Prediction Model

In this section, we propose a visual saliency prediction model based on a semantic segmentation algorithm, where the fixation map is modeled as the foreground (salient object). A semantic segmentation algorithm classifies and labels every pixel in an image into objects (foreground) and background [22]. There are many applications for semantic segmentation, including road segmentation for autonomous driving and cancer cell segmentation for medical diagnosis.

The architecture of the proposed semantic segmentation model is illustrated in Figure 2.2. To obtain a multi-level prediction, each output of the convolution layer (encoder) must be connected directly to the corresponding deconvolution layer (decoder). In general, the task of visual attention uses a combination of low-level and high-level features. In other words, we incorporate multi-layer information together to produce the output saliency map. Low-level features, such as edges, corners, and orientations, are captured by small-level receptive fields, while high-level features, such as semantic information (e.g., object parts or faces), are extracted by high-level receptive fields. Moreover, there are many receptive field sizes, and each corresponds to the layer size. Therefore, based on the advantages of CNNs, we can use small and high receptive fields in down-sampling (e.g., multi-

convolution layers, such as in the VGG-16 network) to create feature maps. Both low- and high-level features are very important for predicting human visual saliency. Thus, our proposed model produces the final saliency map based on the combination of all the outputs of the individual deconvolution operations. Additionally, in our proposed model, we only consider the CNN layers that create feature maps and we exclude the fully connected layers. In addition, the saliency combination block represents the merged multi-layer output saliency predictions (i.e., the prediction average achieves a higher performance compared to that of a single-layer output).

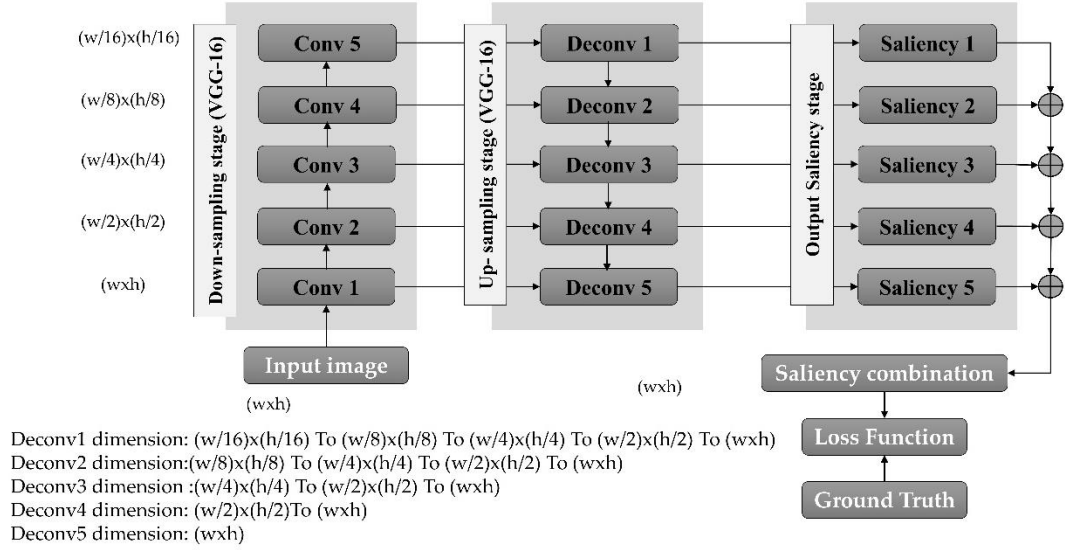


Figure 2.2: Architecture of the proposed model. Note that size of the input image is denoted by $(w \times h)$ where the w is the width, and h is the height in pixels. All saliency maps also have a similar size to that of the input image.

Assume we have an input image, and its feature map is F^{l-1} of the l -th layer and the convolution processes are specified by the weight, W^l . Thus, the output of the feature map can be calculated by:

$$F^l = f_{con}(F^{l-1}; W_{con}^l) = W_{con}^l * F^{l-1}, l = 1 \dots L, \quad (2.4)$$

where F^0 is the input image, the symbol $*$ indicates the convolution operation, and L is the number of layers. The deconvolution operation is the opposite of the convolution operation and it can be run in two directions (forward and backward through of convolution), where it performs the up-sampling operation represented by Equation (2.5):

$$f_{decon}(F; W_{decon}) = W_{decon} \oslash_s F, \quad (2.5)$$

where the \oslash_s is the stride convolution and s is an up-sampling factor. The output operation of the decoder is then given as follows:

$$Y^l = D(F^l; W_{decon}^l), \quad (2.6)$$

where D is the deconvolution operation, and W_{decon}^l is all the kernel weights of the deconvolution layers. Moreover, the total number of weights can be explained by:

$$W = (W_{con}^1, \dots, W_{con}^L, W_{decon}^{l1}, \dots, W_{decon}^{lM}), \quad (2.7)$$

where M is the output prediction maps. Additionally, the loss-function is a Stochastic Gradient Descent with Momentum (SGDM, Equation (2.8)). The objective of this function is to accelerate gradient vectors in the right direction and increase the speed of convergence. In other words, SGDM optimizes the differentiable function and decreases classification errors [19,23]. The loss function can also be defined by Equation (2.8):

$$L(\alpha) = Y \log Hl + (1 - Y) \log(1 - Hl), \quad Y \in \{0,1\}, \quad (2.8)$$

where $L(\alpha)$ is the cross entropy between the predicted probability Hl and the ground truth (GT) labeled Y .

2.3 Materials and Methods

In this section, we describe all the steps for implementing our work, including training, adjusting the parameters of, validating, and testing the model on several available benchmark datasets (TORONTO, MIT300, MIT1003, and DUT-OMRON).

2.3.1 Model Training

The proposed model was trained on a standard dataset (i.e., SALICON) [15]. This dataset consists of a training dataset (10,000 images) and validation dataset (5,000 images) both with ground truth data, and a test dataset (5,000 images) without ground truth data. All the

images are in JPG format, except for the ground truth dataset which is in grayscale PNG format, and all images have a resolution of (640×480). At the beginning of the training, all the weights of the filters were initialized based on the pre-trained network (VGG-16), which has an input image of (224×224) and a Gaussian distribution with a 0.01 standard deviation and zero mean for the weights of each layer [15]. The purpose of using the VGG-16 pre-trained network is to transfer the learned knowledge and reuse it to predict human visual saliency. Additionally, the network parameters were as follows: Initial Learn Rate 0.01, Max Epochs 10, Mini Batch Size 10, and number of iterations 620. The network has been trained on 10,000 images and used selected images from the test datasets for testing (the global accuracy of the proposed model was 96.22%). Moreover, using the loss function (SGDM), the model parameters learned to increase the speed of convergence and to decrease output errors. Figure 2.3 illustrates the training progress produced by the proposed model from the specified training images (SALICON).

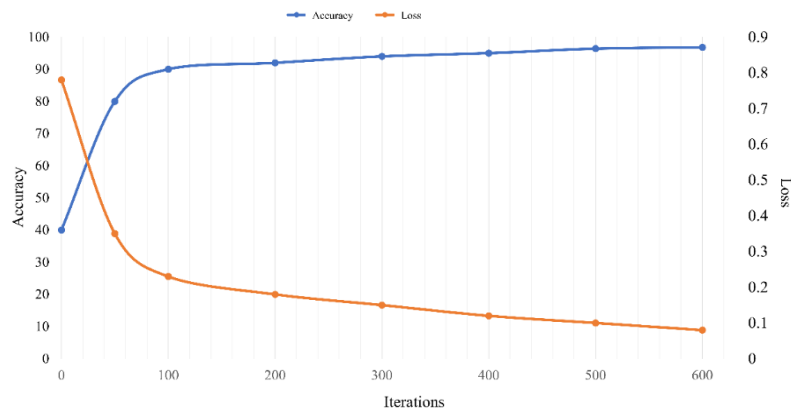




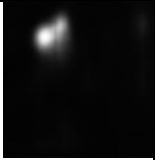


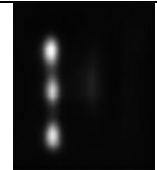
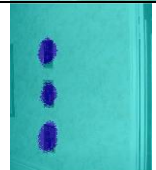
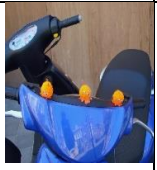
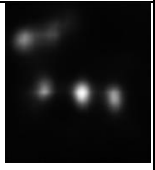




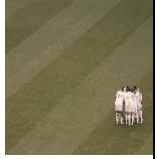
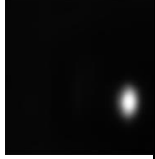
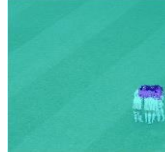



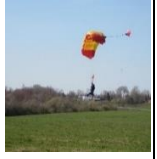
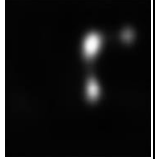




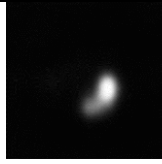


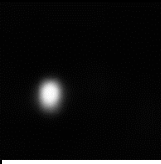



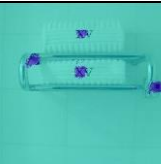

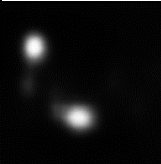



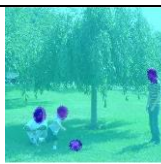

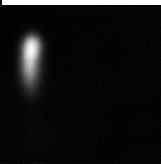


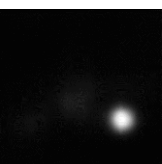


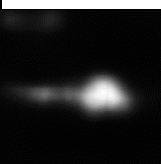

Figure 2.3: Value of validation accuracy and loss as a function of epochs.

2.3.2 Model Testing

This section is for testing the proposed model on several dataset images (test images). As we illustrated in the previous section, the SALICON test images are available without ground truth, thus, we suggested to use other datasets, such as TORONTO, MIT300, MIT1003, and DUT-OMRON datasets for model testing. Figure 2.4 shows the model testing on the selected images. Note that the proposed model has the ability to detect the most salient objects in the scene.

TORONTO			MIT300		
Test Image	Ground Truth	Model Prediction	Test Image	Ground Truth	Model Prediction
					
					
					
					

(a)

MIT300			DUT-OMRON		
Test Image	Ground Truth	Model Prediction	Test Image	Ground Truth	Model Prediction
					
					
					
					

(b)

Figure 2.4: Model testing :(a) TORONTO and MIT300 datasets, and (b) MIT1003 and DUT-OMRON datasets.

2.3.3 Datasets

The proposed model was tested on several well-known datasets, including TORONTO, MIT 300, MIT1003, and DUT-OMRON, which are described below. During the model testing, given an inquiry image, we obtained the saliency map prediction from the last saliency combination layer. The average time to test an image was about 15s.

2.3.3.1 SALICON

SALICON is the largest dataset for visual attention on the popular Microsoft Common Objects in context (MS COCO) image database [15]. It contains 10,000 training images, 5,000 validation images, and 5,000 testing images with a fixed resolution of 480×640 , collected from the Microsoft COCO dataset. This dataset also includes the ground truth data for the training and validation datasets; however, the ground truth data for the test datasets were not available [15].

2.3.3.2 TORONTO

The TORONTO dataset contains 120 colour images with a fixed resolution of 511×681 pixels. This dataset contains both indoor and outdoor environments and was free-viewed by 20 human subjects [17].

2.3.3.3 MIT300

MIT300 is a collection of 300 images that contains the eye movement data of 39 observers. It should be noted that MIT300 is a challenging dataset since its images are highly varied and natural. Saliency maps of all images are withheld and employed by the MIT Saliency Benchmark for model evaluation (http://saliency.mit.edu/results_mit300.html) [24].

2.3.3.4 MIT1003

MIT1003 is a collection of 1,003 images from the Flickr and LabelMe collections. Saliency maps have also been obtained from the eye-tracking data of 15 users. It is the

largest eye fixation dataset, wherein there are 779 landscapes and 228 portraits images that vary in size from 405×405 to 1024×1024 pixels [24].

2.3.3.5 DUT-OMRON

DUT-OMRON contains 5,168 high quality images that were manually selected from more than 140,000 images. Images in this database have one or more salient objects and a relatively complex background [25].

2.3.4 Evaluation Metrics

There are several indices for evaluation metrics to measure the agreement between visual saliency and model prediction. There are also previous studies on saliency metrics, which explain that it's hard to perform a fair comparison to evaluate saliency models by one metric [26]. In general, saliency evaluation indices are divided into location-based and distribution-based metrics. The former type of evaluation considers the saliency map at distinct locations; the latter considers both predicted saliency and human eye fixation maps as continuous distributions. The most well-known location-based indices are the Area Under the ROC curve in two versions of Judd and Borji [22]. Alternatively, the most commonly used distribution-based indices are the Normalized Scanpath Saliency (NSS) and Similarity Metrics (SIM). These indices are described in detail in the following sections [22].

2.3.4.1 Normalized Scanpath Saliency (NSS)

The NSS metric was introduced to the saliency community as a simple correspondence measure between human eye fixation and model prediction. NSS is susceptible to false positives and relative differences in saliency across the image. Given a saliency map S and a binary map of fixation location F , then

$$NSS = \frac{1}{N} \sum_{i=1}^N \bar{S}(i)F(i), \quad (2.9)$$

$$\text{where } N = \sum_i F(i) \quad \text{and} \quad \bar{S} = \frac{S - \mu(S)}{\sigma(S)},$$

where N is the total number of human eye positions and $\sigma(S)$ is the standard deviation.

2.3.4.2 Similarity Metric (SIM)

The similarity metric (SIM) uses the normalized probability distributions of the saliency and human eye fixation maps. SIM is calculated as the sum of the minimum values of each pixel. The similarity between these two maps is calculated as:

$$SIM = \sum_{i=1} \min(\hat{S}(i), \hat{G}(i)), \quad (2.10)$$

where \hat{S} and \hat{G} are the normalized saliency map and the fixation map, respectively. A similarity score between zero and one indicates that the distributions are the same and that they do not overlap.

2.3.4.3 Judd Implementation (AUC-Judd)

The AUC-Judd metric is widely used to evaluate saliency models. The saliency map is treated as a binary classifier to separate positive from negative samples at various thresholds. The true positive (tp) rate is the proportion of the saliency map's values above a certain threshold at fixation locations. The false positive (fp) rate is the proportion of the saliency map's values that occur above the threshold of non-fixated pixels. In this implementation, the thresholds are sampled from the saliency map's values [27,28].

2.3.4.4 Borji Implementation (AUC-Borji)

The AUC-Borji metric uses a uniform random sample of image pixels as negatives and defines the fixation map's (saliency map) values above the threshold of these pixels as false positives. This version of the Area Under ROC curve measurement is based on Ali Borji's code. The saliency map is treated as a binary classifier to separate positive from negative samples at various thresholds. The true positive (TP) rate is the proportion of the saliency map's values above the threshold of fixation locations. The false positive (FP) rate is the proportion of the saliency map's values that occur above the threshold sampled from random pixels (as many samples as fixations, sampled uniformly from all image pixels). In this implementation, threshold values are sampled at a fixed step size [29].

2.3.4.5 Semantic Segmentation Metrics

These metrics are used to evaluate the prediction results against the ground truth data. In this study two different semantic segmentation metrics are used, which Global Accuracy

and Weighted Intersection over Union (WeightedIoU). Specifically, the Global Accuracy is the ratio of correctly classified pixels, regardless of class, to the total number of pixels and the WeightedIoU is the average IoU of all classes, weighted by the number of pixels in the class, wherein the MeanIoU is the average IoU score of all classes in that particular image [27,30].

2.4 Experimental Results

2.4.1 Quantitative Comparison of the Proposed Model with Other State-of-the-Art Models

To evaluate the efficiency of the proposed model, we compared it to six state-of-the-art models. We selected four dataset benchmarks (TORONTO ,MIT300, MIT1003, and DUT-OMRON) for comparison of the quantitative results. These results are reported in Table 2.1, Table 2.2, Table 2.3, and Table 2.4, respectively.

Table 2.1 shows that, with the TORONTO dataset, the proposed model outperforms the other six models in terms of the NSS, AUC-Judd, and AUC-Borji metrics; however, in terms of the SIM (similarity) metric, the DVA algorithm [19] has the best results. This is because SIM metric is better suited for non-binary classifiers. However, the proposed algorithm is a binary classifier. The other metrics used in the study (NSS, AUC-Judd, AUC-Borji) are all binary classifier metrics. From Table 2.2, one can see similar results for the MIT300 dataset as those for the TORONTO dataset, except for the AUC-Borji metric, where the GBVS and Judd models perform slightly better than the proposed model.

Table 2.3 illustrates that for the MIT1003 dataset, the proposed model again outperforms the other six models in terms of the NSS and AUC-Judd metrics; however, in terms of the other two metrics, the DVA model provides the best performance. From Table 2.4, one can see that for the DUT-OMRON dataset the proposed model outperforms the other six models only in terms of the AUC-Judd metric and the DVA model provides the best performance in terms of the other three metrics. Overall, for all four investigated datasets, the proposed model provides the highest AUC-Judd metric.

Table 2.5 explains the evaluation metrics obtained from the proposed model. Specifically, the highest and lowest Global Accuracies are obtained when the model was tested on the TORONTO dataset (global accuracy of 96.22%), and the MIT300 dataset (global accuracy of 94.13%), respectively.

Table 2.1: Comparison of the quantitative scores of several models on TORONTO [24] dataset. Note, the bold values are the best scores.

Model	NSS	SIM	AUC-Judd	AUC-Borji
ITTI [31]	1.30	0.45	0.80	0.80
AIM [16]	0.84	0.36	0.76	0.75
Judd Model [27]	1.15	0.40	0.78	0.77
GBVS [24]	1.52	0.49	0.83	0.83
Mr-CNN [32]	1.41	0.47	0.80	0.79
DVA [19]	2.12	0.58	0.86	0.86
Proposed Model	3.00	0.42	0.91	0.87

Note. Humans baseline [22] 3.29 1.00 0.92 0.88

Table 2.2: Comparison of the quantitative scores of several models on MIT300 [24] dataset.

Model	NSS	SIM	AUC-Judd	AUC-Borji
ITTI	0.97	0.44	0.75	0.74
AIM	0.79	0.40	0.77	0.75
Judd Model	1.18	0.42	0.81	0.80
GBVS	1.24	0.48	0.81	0.80
Mr-CNN	1.13	0.45	0.77	0.76
DVA	1.98	0.58	0.85	0.78
Proposed Model	2.43	0.51	0.87	0.80

Table 2.3: Comparison of the quantitative scores of several models on MIT1003 [24] dataset.

Model	NSS	SIM	AUC-Judd	AUC-Borji
ITTI	1.10	0.32	0.77	0.76
AIM	0.82	0.27	0.79	0.76
Judd Model	1.18	0.42	0.81	0.80
GBVS	1.38	0.36	0.83	0.81
Mr-CNN	1.36	0.35	0.80	0.77
DVA	2.38	0.50	0.87	0.85
Proposed Model	2.39	0.42	0.87	0.80

Table 2.4: Comparison of the quantitative scores of several models on DUT-OMRON [24] dataset.


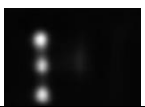
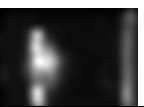


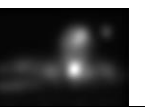

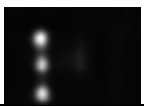
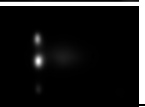

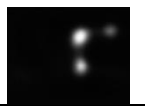
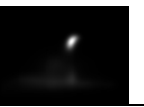
Model	NSS	SIM	AUC-Judd	AUC-Borji
ITTI	3.09	0.53	0.83	0.83
AIM	1.05	0.32	0.77	0.75
GBVS	1.71	0.43	0.87	0.85
DVA	3.09	0.53	0.91	0.86
Proposed Model	2.50	0.49	0.91	0.84

Table 2.5: Model predication results (i.e., Global accuracy) on several datasets (TORONTO, MIT300, MIT1003, and DUT-OMRON).

Datasets	GlobalAccuracy	WeightedIoU
TORONTO	0.96227	0.94375
MIT300	0.94131	0.91924
MIT1003	0.94862	0.92638
DUT-OMRON	0.94484	0.92605

2.4.2 Qualitative Comparison of the Proposed Model with Other State-of-the-Art Models

We first qualitatively tested the proposed model on the SALICON dataset; then, we evaluated the model on the TORONTO, MIT300, MIT1003, and DUT-OMRON datasets. Figures 2.5 illustrates the saliency map results obtained when the proposed model and five other state-of-the-art models are applied to sample images drawn from the studied datasets. From this figure, one can see that the proposed model is capable of predicting most of the salient objects in the given images.

-----	TORONTO			MIT300		
Model	Test image	Ground Truth	Model prediction	Test image	Ground Truth	Model prediction
ITTI						
FES [12]						


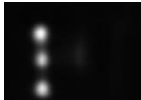
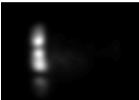




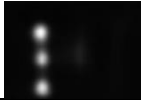
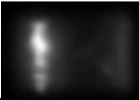

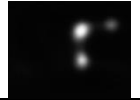



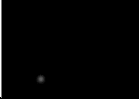








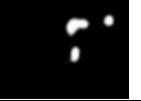

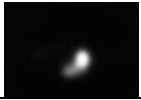
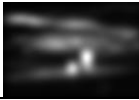





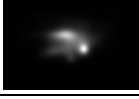


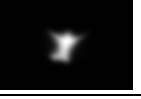


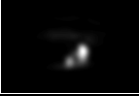




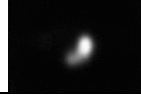



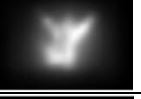


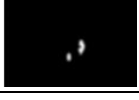





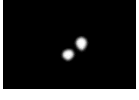



CovSal [30]						
GBVS						
DVA						
Proposed Model						
-----	MIT1003			DUT-OMRON		
ITTI						
FES						
CovSal						
GBVS						
DVA						
Proposed Model						

Figure 2.5: The saliency maps obtained from the proposed model and five other state-of-the-art models for a sample image from the TORONTO, MIT300, MIT1003, and DUT-OMRON datasets.

2.5 Conclusion

In this study, a deep learning model has been proposed to predict visual saliency on images. This work uses a deep network with five encoders and five decoders (convolution and deconvolution) and the semantic segmentation approach to predict human visual saliency. The proposed model generates a sequence of features at the multi-stage level to produce a saliency map. The experimental results obtained from the analysis of four benchmark datasets illustrate the superior prediction capability of the proposed model with respect to other state-of-the-art methods. Additionally, the proposed model achieved an accuracy of more than 94% for all datasets, although the highest performance (i.e., 96%) was obtained from the TORONTO dataset. Additionally, in the training stage, the increased number of training images will increase the prediction accuracy of the proposed model; however, the model requires a larger memory.

In the future, we will focus on how to collect a new dataset, creating its ground truth data (e.g., data augmentation method), and designing new models with improved evaluation metrics. Importantly, it is possible to use the model presented herein to facilitate other tasks, such as salient object detection, scene classification, and object detection. Moreover, this work provides the basis to develop new models, able to learn from high-level understanding; for example, they will be able to detect the most interesting part of the image (e.g., a human face) and the most prominent person in the scene.

References

1. Sun, Y.; Fisher, R. Object-based visual attention for computer vision. *Artif. Intell.* **2003**, *146*, 77–123.
2. Koch, C.; Ullman, S. Shifts in selective visual attention: Towards the underlying neural circuitry. In *Matters of Intelligence*; Springer: Dordrecht, The Netherlands, 1987; pp. 115–141.
3. Wang, K.; Wang, S.; Ji, Q. Deep eye fixation map learning for calibration-free eye gaze tracking. In Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, Charleston, SC, USA, 14–17 March 2016; pp. 47–55.
4. Borji, A. Boosting bottom-up and top-down visual features for saliency estimation. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 438–445.
5. Zhu, W.; Deng, H. Monocular free-head 3D gaze tracking with deep learning and geometry constraints. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3143–3152.
6. Kanan, C.; Tong, M.H.; Zhang, L.; Cottrell, G.W. SUN: Top-down saliency using natural statistics. *Vis. Cogn.* **2009**, *17*, 979–1003.
7. Hickson, S.; Dufour, N.; Sud, A.; Kwatra, V.; Essa, I. Eyemotion: Classifying facial expressions in VR using eye-tracking cameras. In Proceedings of the 2019 IEEE

- Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 1626–1635.
8. Zhao, R.; Ouyang, W.; Li, H.; Wang, X. Saliency detection by multi-context deep learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1265–1274.
 9. Recasens, A.; Vondrick, C.; Khosla, A.; Torralba, A. Following gaze in video. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1435–1443.
 10. Wang, C.; Shi, F.; Xia, S.; Chai, J. Realtime 3D eye gaze animation using a single RGB camera. *ACM Trans. Graph.* **2016**, *35*, 118.
 11. Cornia, M.; Baraldi, L.; Serra, G.; Cucchiara, R. Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Trans. Multimed. Comput. Commun. Appl.* **2018**, *14*, 48.
 12. Naqvi, R.; Arsalan, M.; Batchuluun, G.; Yoon, H.; Park, K. Deep learning-based gaze detection system for automobile drivers using a NIR camera sensor. *Sensors* **2018**, *18*, 456.
 13. Rezaee, M.; Mahdianpari, M.; Zhang, Y.; Salehi, B. Deep Convolutional Neural Network for Complex Wetland Classification Using Optical Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3030–3039.

14. Krafska, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; Torralba, A. Eye tracking for everyone. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2176–2184.
15. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
16. Kruthiventi, S.S.S.; Gudisa, V.; Dholakiya, J.H.; Venkatesh Babu, R. Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016; pp. 5781–5790.
17. Pan, J.; Sayrol, E.; Giro-i-Nieto, X.; McGuinness, K.; O’Connor, N.E. Shallow and deep convolutional networks for saliency prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 598–606.
18. Mahdianpari, M.; Salehi, B.; Mohammadimanesh, F.; Motagh, M. Random forest wetland classification using ALOS-2 L-band, RADARSAT-2 C-band, and TerraSAR-X imagery. *ISPRS J. Photogramm. Remote Sens.* **2017**, *130*, 13–31.

19. Liu, N.; Han, J.; Liu, T.; Li, X. Learning to predict eye fixations via multiresolution convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *29*, 392–404.
20. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
21. Mahdianpari, M.; Salehi, B.; Rezaee, M.; Mohammadimanesh, F.; Zhang, Y. Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery. *Remote Sens.* **2018**, *10*, 1119.
22. Jiang, M.; Huang, S.; Duan, J.; Zhao, Q. Salicon: Saliency in context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1072–1080.
23. Judd, T.; Ehinger, K.; Durand, F.; Torralba, A. Learning to predict where humans look. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 2106–2113.
24. Bruce, N.; Tsotsos, J. Saliency based on information maximization. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 4–7 December 2006; pp. 155–162.
25. Li, Y.; Hou, X.; Koch, C.; Rehg, J.M.; Yuille, A.L. The secrets of salient object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 280–287.

26. Wang, W.; Shen, J. Deep visual attention prediction. *IEEE Trans. Image Process.* **2017**, *27*, 2368–2378.
27. Mohammadimanesh, F.; Salehi, B.; Mahdianpari, M.; Gill, E.; Molinier, M. A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem. *ISPRS J. Photogramm. Remote Sens.* **2019**, *151*, 223–236.
28. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
29. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
30. Qian, N. On the momentum term in gradient descent learning algorithms. *Neural Netw.* **1999**, *12*, 145–151.
31. Judd, T.; Durand, F.; Torralba, A. A Benchmark of Computational Models of Saliency to Predict Human Fixations, 2012. Available online: <http://hdl.handle.net/1721.1/68590> (accessed on 9 August 2019).
32. Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; Yang, M.-H. Saliency detection via graph-based manifold ranking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3166–3173.

33. Wang, W.; Shen, J.; Yang, R.; Porikli, F. Saliency-aware video object segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 20–33.
34. Harel, J.; Koch, C.; Perona, P. Graph-based visual saliency. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 3–6 December 2007; pp. 545–552.
35. Borji, A.; Tavakoli, H.R.; Sihite, D.N.; Itti, L. Analysis of scores, datasets, and models in visual saliency prediction. In Proceedings of the IEEE international Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 921–928.
36. Tavakoli, H.R.; Rahtu, E.; Heikkilä, J. Fast and efficient saliency detection using sparse sampling and kernel density estimation. In Proceedings of the Scandinavian Conference on Image Analysis, Ystad, Sweden, 23–25 May, 2011; pp. 666–675.
37. Erdem, E.; Erdem, A. Visual saliency estimation by nonlinearly integrating features using region covariances. *J. Vis.* **2013**, *13*, 11.
38. Bruce, N.D.B.; Tsotsos, J.K. Saliency, attention, and visual search: An information theoretic approach. *J. Vis.* **2009**, *9*, 5.
39. Csurka, G.; Larlus, D.; Perronnin, F.; Meylan, F. What is a good evaluation measure for semantic segmentation? In Proceedings of the 24th British Machine Vision Conference (BMVC), Bristol, UK, 9–13 September 2013; Volume 27, p. 2013.

Chapter 3. Performance Evaluation of Pre-Trained CNN Models for Visual Saliency Prediction

Abstract

The Human Visual System (HVS) has the ability to focus on specific parts of a scene, rather than the whole scene. This phenomenon is one of the most active research topics in the computer vision and neuroscience fields. Recently, deep learning models have been used for visual saliency prediction. In this work, we investigate the performance of five state-of-the-art deep neural networks (VGG-16, ResNet-50, Xception, InceptionResNet-v2, and MobileNet-v2) for the task of visual saliency prediction. In this work, we train five deep learning models over the SALICON dataset and then use the trained models to predict visual saliency maps using four standard datasets, namely: TORONTO, MIT300, MIT1003, and DUT-OMRON. The results indicate that the ResNet-50 model outperforms the other four and provides a visual saliency map that is very close to human performance.

3.1 Introduction

The Human Visual System (HVS) processes certain parts of a visual scene rather than the entire image. This is called Human Visual Attention (HVA), also referred to as visual saliency prediction. Visual saliency prediction is also beneficial for other applications in the computer vision field, including salient object detection [1], image retrieval [2], multiresolution imaging [3], and scene classification [4-6].

Over the past few years, several models have been proposed to tackle the problem of visual saliency prediction. The most common form used to predict visual saliency is a saliency map. A saliency map predicts the probability that each pixel in an image will attract human attention. To generate a saliency map, the salient points in the image are collected and convolved with a Gaussian filter [7]. Generally, saliency maps smooth the image, making it more meaningful and easier to analyze. Overall, HVA is sorted into bottom-up and top-down approaches. The first approach uses low-level features, including intensity, color, texture, and edge orientation [8, 9]. The second approach is a top-down method, which is task-driven and requires an explicit understanding of the context of the visual scene. On the other hand, it depends on the features of the object of interest [10, 11].

Deep Convolutional Neural Networks (CNNs) have been used in the field of visual attention because they can extract robust features and achieve superior performance compared with other state-of-the-art methods. For instance, Fully Convolutional Neural networks (FCNs) have been recently proposed to solve the problem of visual saliency prediction [12]. FCNs are beneficial because they have the same architecture as CNNs but do not contain fully connected layers.

In this chapter, we use FCN architecture for semantic segmentation based on five deep learning models, namely VGG-16, ResNet-50, Xception, InceptionResNet-v2, and MobileNet-v2 to solve the problem of visual saliency prediction. These models were first trained on the SALICON dataset, and the trained models were then evaluated over four datasets, including TORONTO, MIT300, MIT1003, and DUT-OMRON.

The remainder of this chapter is organized as follows: Section 3.2 describes the overview of semantic segmentation and pre-trained models. Section 3.3 explains the visual and numerical experimental results. Finally, this work is concluded in Section 3.4.

3.2 Materials and Methods

3.2.1 Semantic Segmentation

Semantic segmentation plays an important role in image understanding and is essential for image analysis tasks. In semantic segmentation, each region or pixel is labeled with a set of classes as backgrounds and foregrounds. Deep neural networks are commonly used as effective techniques for semantic segmentation. In this section, we briefly introduce the five state-of-the-art deep learning models used in this work.

3.2.2 Pre-Trained Models

Pre-trained models are those trained over a large benchmark dataset to classify images from different classes. These pre-trained models can be used for other applications, such as semantic segmentation using a technique called transfer learning. Using pre-trained models finely tuned by transfer learning is beneficial when only a small training dataset is available. The models used in this work have all been trained over huge datasets (e.g., ImageNet dataset [13]). They can classify images into 1000 classes, such as keyboard, mouse, pencil, and many types of animals. Consequently, the models have learned rich feature representations of a wide range of images, subsequently adjusting the parameters for visual saliency dataset.

3.2.2.1 VGG-16 Model

The VGG-16 network was developed by Simonyan and Zisserman in the 2014 ILSVRC competition [14]. Generally, the VGG-16 network contains thirteen convolution layers, five pooling layers, and three fully connected layers. The VGG-16 model can classify images into 1000 object classes and has an image input size of 224×224 .

3.2.2.2 ResNet-50 Model

This model is a Convolutional Neural Network that was trained over more than a million images from the ImageNet database [13]. This model has 50 layers and can classify images into 1000 object categories [15].

3.2.2.3 Xception Model

Xception is a CNN introduced by Francois Chollet at the Conference on Computer Vision and Pattern Recognition (CVPR) in 2017. This model has 18 layers and an image input size of 299×299 [16].

3.2.2.4 InceptionResNet-V2 Model

This type of the network was built by integrating two deep CNNs, namely ResNet [15] and inception models [17]. This model has 164 layers and an image input size of 299×299 . This model also has the ability to classify images into 1000 object classes.

3.2.2.5 MobileNet-V2 Model

This model is a CNN that was trained over more than a million images from the ImageNet database. This model is 54 layers deep and has an image input size of 224×224 [18]. This can also classify images into 1000 object classes.

3.2.3 Pre-Trained Models Specification

Table 3.1 presents some of the parameters of the pre-trained models used in this study. These parameters are important as they affect the performance and complexity (training and testing time) of each model.

Table 3.1: Pre-Trained Model Parameters. Note, **L**: layers, **M**: Million.

Pre-Trained Model	Depth (L)	Image	Number of Parameters (M)
VGG-16	16	224×224	138
ResNet-50	50	224×224	25.6
Xception	71	299×299	22.9
InceptionResNet-v2	164	299×299	55.9
MobileNet-v2	53	224×224	3.5

3.2.4 Datasets

The models were trained using the SALICON dataset and then tested on four other standard datasets, including TORONTO, MIT 300, MIT1003, and DUT-OMRON described below:

3.2.4.1 SALICON Dataset

SALICON is the largest dataset for visual attention in the popular Microsoft Common Objects in Context (MS COCO) image database [19]. It contains 10,000 training images, 5000 validation images, and 5000 testing images, with a fixed resolution of 480×640 pixels.

3.2.4.2 TORONTO Dataset

The TORONTO dataset contains 120 colour images with a fixed resolution of 511×681 pixels. This dataset contains both indoor and outdoor environments and was free viewed by 20 human subjects [20].

3.2.4.3 MIT300 Dataset

The MIT300 dataset has 300 natural images, and their saliency maps were generated from the eye-tracking data of 39 users who free-viewed these images. This dataset is a challenging dataset since its images are highly varied and natural [4].

3.2.4.4 MIT1003 Dataset

MIT1003 is a collection of 1003 images from the Flickr and LabelMe collections. Saliency maps were also obtained from the eye-tracking data of 15 users. It is the largest eye fixation dataset, wherein there are 779 landscapes and 228 portraits images that vary in size from 405×405 to 1024×1024 pixels [4].

3.2.4.5 DUT-OMRON Dataset

DUT-OMRON has 5,168 high quality images. The largest height or width of this dataset is 400 pixels. Each image was free viewed by five subjects. This dataset was manually selected from more than 140,000 images. There is more than one salient object in this dataset, and each image has a more complex background [21].

3.2.5 Evaluation Metrics

There are several methods to evaluate the agreement between human eye fixation and model prediction. In this work, we utilize the following four evaluation metrics for performance assessment.

3.2.5.1 Normalized Scanpath Saliency (NSS)

The NSS metric was introduced to the saliency community as a simple correspondence measure between ground truth data and model prediction [20].

3.2.5.2 Similarity Metric (SIM)

SIM was presented to the saliency community as a simple correspondence measure between saliency maps and ground truth data, computed as the average normalized saliency at fixated locations [20].

3.2.5.3 AUC-Borji

This version of the Area Under ROC Curve (AUC)-measure was developed by A. Borji. This metric uses a uniform random sample of image pixels as negatives and defines the fixation (saliency) map's values above the threshold of these pixels as false positives [20].

3.2.5.4 AUC-Judd

This version of the AUC-measure was developed by T. Judd. The AUC-Judd metric is widely used for evaluating saliency models. The saliency map is treated as a binary classifier to separate positive from negative samples at various thresholds [20].

3.3 Experimental Results

3.3.1 Models Training

Since this work is based on the transfer learning method (fine-tuning), a large training dataset is not required. Here, the five pre-trained models presented in Section 3.2.2 were trained over a standard dataset (i.e., SALICON). In the training stage, 1000 images were employed. Table 3.2 illustrates the important parameters (i.e., training global accuracy, validation accuracy, and training time) provided after the training process. The investigated models were trained using minimization of the cross-entropy loss function and stochastic gradient descent with a momentum (SGDM) optimizer. The mini-batch size was 10, the number of epochs was 25, and the learning rate was set at 0.001.

Table 3.2: The most important parameters through the training stage of all pre-trained models. Note, **AC**: Accuracy, **LO**: Loss.

-----	Training		Validation		Training Time (min)	Testing Accuracy (%)
Pre-Trained Model	AC (%)	LO (%)	AC (%)	LO (%)		
VGG-16	94.3	0.15	92.9	0.17	602	92.19
ResNet-50	99.4	0.02	95.2	0.25	1602	95.10
Xception	99.6	0.01	95.2	0.33	2239	94.99
InceptionResNet-v2	99.6	0.01	95.5	0.30	2683	95.28
MobileNet-v2	99.0	0.03	95.4	0.22	1118	95.22

3.3.2 Visual Results

Fig 3.1 depicts the visual saliency maps predicted by the state-of-the-art pre-trained models discussed in Section 3.2.2 and the datasets presented in Section 3.2.4. From this figure, one can see that both VGG-16 and ResNet-50 predict visual saliency maps very close to the ground truth. One can further see that the poorest result was obtained from the InceptionResNet-v2 model. Among the four datasets, the image drawn from the TORONTO dataset provides better results than those drawn from the other three datasets.


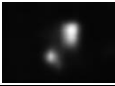



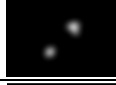


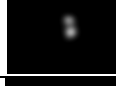

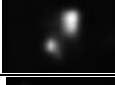
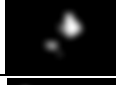

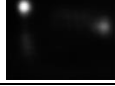





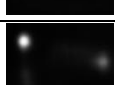


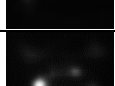


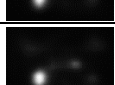


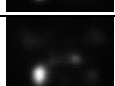
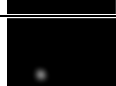

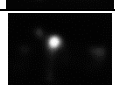
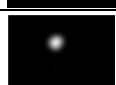

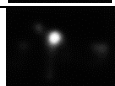
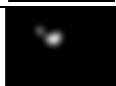
Pre-Trained Model Dataset	Input Image	Ground Truth	Model Prediction
VGG-16 TORONTO			
ResNet-50 TORONTO			
Xception TORONTO			
MobileNet-v2 TORONTO			
VGG-16 MIT300			
MobileNet-v2 MIT300			
InceptionResNet-v2 MIT300			
ResNet-50 MIT1003			
Xception MIT1003			
InceptionResNet-v2 MIT1003			
VGG-16 DUT-OMRON			
ResNet-50 DUT-OMRON			

Figure 3.1: Visual saliency maps predicted by state-of-the-art pre-trained models for images drawn from the TORONTO, MIT300, MIT1003, and DUT-OMRON datasets.

3.3.3 Numerical Results

To give more insight on the performance of the five investigated models, we provide numerical results in this Section. For comparison purpose, we use the metrics presented in Section 3.2.5, i.e., SIM, NSS, AUC-Judd, and AUC-Borji.

Table 3.3 presents the SIM and NSS metrics computed for all four datasets. From the left part of this table, one can see that the VGG-16 model provides the highest SIM value; however, for the other three datasets, the SIM metric is mostly lower than the other four models. On the other hand, ResNet-50 performs consistently well for all four datasets. From the right side of Table 3.3, one can see that all models provide reasonable NSS values for all datasets; however, the lowest NSS is obtained using InceptionResnet-v2 model which has even more layers than the other four models and this is mainly due to the overfitting problem which usually observed in very deep models. Table 3.4 presents the AUC-Judd and AUC-Borji metrics for all four datasets. From this table, one can see that the ResNet-50 model consistently performs well and, on average, outperforms the other models. The poorest AUC metrics were obtained using the InceptionResNet-v2 model.

Table 3.3: SIM and NSS metrics obtained from the state-of-the-art pre-trained models for the **A**: TORONTO, **B**: MIT300, **C**: MIT1003, and **D**: DUT-OMRON datasets. Note: Bolded values represent the best values.

-----	SIM				NSS			
Pre-Trained Model	A	B	C	D	A	B	C	D
VGG-16	3.00	0.97	1.13	0.64	0.42	0.32	0.36	0.47
ResNet-50	1.48	1.49	1.79	1.53	0.45	0.35	0.30	0.41
Xception	1.51	1.67	2.21	1.30	0.24	0.12	0.21	0.28
InceptionResNet-v2	1.10	1.89	1.91	1.35	0.28	0.14	0.18	0.28
MobileNet-v2	1.42	1.14	1.09	1.35	0.37	0.39	0.37	0.33

Table 3.4: AUC-Judd and AUC-Borji metrics obtained from the state-of-the-art pre-trained models for **A**: TORONTO, **B**: MIT300, **C**: MIT1003, and **D**: DUT-OMRON datasets. Note: Bolded values represent the best values.

-----	AUC-Judd				AUC-Borji			
Pre-Trained Model	A	B	C	D	A	B	C	D
VGG-16	0.91	0.86	0.87	0.76	0.87	0.66	0.68	0.59
ResNet-50	0.92	0.90	0.90	0.91	0.73	0.74	0.78	0.74
Xception	0.88	0.89	0.86	0.84	0.75	0.76	0.82	0.70
InceptionResNet-v2	0.81	0.92	0.85	0.82	0.68	0.83	0.78	0.69
MoileNet-v2	0.89	0.91	0.86	0.87	0.72	0.71	0.68	0.71

3.4 Conclusion

In this work, we assessed the performance of five state-of-the-art deep learning models (i.e., VGG-16, ResNet-50, Xception, InceptionResNet-v2, and MobileNet-v2) for visual saliency prediction. These models were trained using the SALICON dataset and then tested over four other standard datasets (i.e., TORONTO, MIT300, MIT1003 and DUT-OMRON). ResNet-50 outperformed the other models and its saliency maps very closely predicts the ground truth data. For instance, the AUC-Judd metric averaged over all four datasets was 0.91. The poorest performance was observed from InceptionResNet-v2 model and was likely caused by overfitting due to the large number of layers in this model.

References

1. Liu, N. and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
2. Huang, J., et al., Integrating visual saliency and consistency for re-ranking image search results. IEEE Transactions on Multimedia, 2011. **13**(4): p. 653-661.
3. Lu, X. and X. Li, Multiresolution imaging. IEEE transactions on cybernetics, 2013. **44**(1): p. 149-160.
4. Cheng, G., et al., Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 2015. **53**(8): p. 4238-4249.
5. Lu, X., X. Li, and L. Mou, Semi-supervised multitask learning for scene recognition. IEEE transactions on cybernetics, 2014. **45**(9): p. 1967-1976.
6. Yao, X., et al., Semantic annotation of high-resolution satellite images via weakly supervised learning. IEEE Transactions on Geoscience and Remote Sensing, 2016. **54**(6): p. 3660-3671.
7. Pan, J., et al., Salgan: Visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081, 2017.
8. Gao, D., V. Mahadevan, and N. Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. in Advances in neural information processing systems. 2008.

9. Le Meur, O., et al., A coherent computational approach to model bottom-up visual attention. *IEEE transactions on pattern analysis and machine intelligence*, 2006. **28**(5): p. 802-817.
10. Gao, D., S. Han, and N. Vasconcelos, Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009. **31**(6): p. 989-1005.
11. Kanan, C., et al., SUN: Top-down saliency using natural statistics. *Visual cognition*, 2009. **17**(6-7): p. 979-1003.
12. Long, J., E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
13. Krizhevsky, A., I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. in *Advances in neural information processing systems*. 2012.
14. Simonyan, K. and A. Zisserman, Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
15. He, K., et al. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
16. Chollet, F. Xception: Deep learning with depthwise separable convolutions. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

17. Szegedy, C., et al. Going deeper with convolutions. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
18. Zhu, M.S.A.H.M. and A.Z.L.-C. Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. in C]/The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.
19. Jiang, M., et al. Salicon: Saliency in context. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
20. Ghariba, B., M.S. Shehata, and P. McGuire, Visual Saliency Prediction Based on Deep Learning. Information, 2019. **10**(8): p. 257.
21. Riche, N., et al. Saliency and human fixations: State-of-the-art and study of comparison metrics. in Proceedings of the IEEE international conference on computer vision. 2013.

Chapter 4. A Novel Fully Convolutional Network for Visual Saliency

Prediction

Abstract

Human Visual System (HVS) has the ability to pay visual attention, which is one of the many functions of the HVS. Despite the many advancements being made in visual saliency prediction, there continues to be room for improvement. Deep learning has recently been used to deal with this task. This work proposes a novel deep learning model based on a Fully Convolutional Network (FCN) architecture. The proposed model is trained in an end-to-end style and designed to predict visual saliency. The model is based on the encoder-decoder structure and includes two types of modules. The first has three stages of inception modules to improve multi-scale derivation and enhance contextual information. The second module includes one stage of the residual module to provide a more accurate recovery of information and simplify optimization. The entire proposed model is fully training style from scratch to extract distinguishing features and to use a data augmentation technique to create variations in the images. The proposed model is evaluated using several benchmark datasets, such as MIT300, MIT1003, TORONTO, and DUT-OMRON. The quantitative and qualitative experiment analyses demonstrate that the proposed model achieves superior performance for predicting visual saliency.

4.1 Introduction

A Human Visual System (HVS) processes a part of the visual scene instead of the whole scene. This phenomenon is called Human Visual Attention (HVA), also referred to as visual saliency prediction, which is an important research area in the field of computer vision. HVA is also known as human eye fixation prediction, visual saliency prediction, or saliency map detection. Visual saliency prediction is also beneficial for other applications in the computer vision field, including salient object detection [1], image retrieval [2], multiresolution imaging [3], and scene classification [4-6].

Many models have been developed to predict visual saliency, the most popular being the saliency map. Saliency maps describe the probability that each image pixel will attract human attention. In other words, saliency maps are images that display the unique qualities of each pixel in a given image [7]. To produce a saliency map, the salient points in the image are collected and convolved with a Gaussian filter [7]. The probability that each pixel in the image will attract human attention is represented by a heat map or gray-scale image. Notably, saliency maps smooth the image, making it more meaningful and easier to analyze. This is useful for condition image captioning architecture because it indicates what is salient and what is not [8].

To evaluate the saliency map, human eye fixation data in free viewing is used because there is a direct link between human eye movement and visual attention [8]. Generally, HVA runs on two approaches. The first is a bottom-up approach which utilizes low-level features, including intensity, color, edge orientation, and texture [9, 10]. Such an approach

attempts to decide regions that show obvious characteristics of their surroundings. The second is a top-down approach, which is task-driven and requires an explicit understanding of the context of the visual scene. Moreover, it depends on the features of the object of interest [11, 12].

The deep Convolutional Neural Network (CNN) is the most widely utilized deep learning method for image processing applications. Specifically, CNN is capable of extracting discriminant visual features (e.g., 2-D spatial features) by applying a hierarchy of convolutional filters using multiple nonlinear transformations. Studies have also used Convolutional Neural Networks (CNNs) for studying saliency map detection to confirm the importance of end-to-end task learning and automatic feature extraction [13-17]. The deep CNN model achieves an even higher classification accuracy. For example, deep learning techniques have achieved superior results in multiple tasks, such as driverless car, scene classification, object (e.g., vehicle) detection, image classification, and semantic segmentation. However, deep learning architecture requires sufficient training data for superior performance on several sets of visual tasks, such as local image detection [18], global image classification[19], and semantic segmentation [20].

Although several deep learning models have been proposed to solve the problem of saliency prediction, and those models provide good performance, those models essentially were proposed for object recognition and then fine-tuned for saliency prediction. Consequently, the pixel-based classification for the visual attention task remains challenging. This highlights the necessity of designing a novel FCN model specifically for

the task of saliency prediction. In addition, our proposed model is designed for training from scratch. Therefore, we added some modules (e.g., three inception modules and residual modules) to improve the model performance.

The inception module is useful since there are benefits from filters with different sizes in one layer, which contribute to multi-scale inference and enhance contextual information. This highlights the necessity of combining feature maps at different resolution to extract useful information. In addition, residual module recovers more accurate information and simplifies optimization, while avoiding the vanishing gradient problem. Moreover, the residual module decreases the number of parameters by dropping several layers in the deep learning model. Therefore, this prevents the overfitting of the proposed model.

In this study, we utilized an encoder-decoder structure based on the Fully Convolutional Network (FCN) architecture to address the problem of bottom-up visual attention in visual saliency predication. FCN has the same architecture as the CNN network, but unlike CNN it does not contain any fully connected layers. FCNs are also powerful visual models that generate high-level features from low-level features to produce hierarchies. Moreover, FCN utilizes multi-layer information and addresses pixel-based classification tasks using an end-to-end style [20]. In addition, the proposed model also includes both inception and residual modules to improve multi-scale inference and the recovery of more accurate information, respectively.

This study proposes a new model based on an encoder-decoder structure (i.e., FCN) to improve the performance of visual saliency prediction. The specific contributions of this work are as follows:

(1) A new model of FCN architecture for visual saliency prediction that uses two types of modules is proposed. The first module contains three stages of inception modules, improves the multi-scale inference, and performs contextual information. The second module contains one stage from the residual module and also recovers more accurate information and simplifies optimization, while avoiding the vanishing gradient problem.

(2) Four well-known datasets, including TORONTO, MIT300, MIT1003, and DUT-OMRON, were used to evaluate the proposed model. The experiments demonstrate that the proposed model achieves results comparable or superior to those of other state-of-the-art models.

The remainder of this work is organized as follows: First, the related work is explained in Section 4.2. The Proposed model is described in more detail in Section 4.3 and the experimental results for the proposed model are discussed in Section 4.4. Section 4.5 presents the quantitative and qualitative experimental results obtained from the four datasets. Finally, the results are summarized, and possible future uses, and applications of the proposed model are explored in Section 4.6.

4.2 Related work

Visual saliency prediction has received attention from computer vision researchers for many years. The earliest computational model was introduced by Koch and Ullman [19], which inspired the work of Itti et al. [21]. This model combines low level features at multiple scales to generate saliency maps. Subsequently, many models have been proposed in this way and in the last several decades there has been renewed interest in visual saliency detection [22-32]. Most of this work has been focused on how to detect visual saliency in an image/video using different methods [33-35].

Most conventional attention models are based on a bottom-up strategy. These contain three important steps to detect visual saliency: feature extraction, saliency extraction, and saliency combination. Salient regions in the visual scene are first extracted from their surroundings through hand-crafted low-level features (e.g., intensity, color, edge orientation, and texture), and center-surround contrast is widely used for generating saliency. The saliency may also be produced by the relative difference between the region and its local surroundings [21], [36], [37]. The last step for saliency detection combines several features to generate the saliency map.

In the last few years, many visual saliency models have been introduced for object recognition. Deep-learning models achieved better performance compared to non-deep learning models. The first proposed model, Deep Neural Networks (DNN) [17], was trained from scratch to predict saliency. Subsequent models were based on pre-trained models. For example, the DeepGaze I model [38] was the first to be trained on a pre-

trained model (AlexNet [19] trained on ImageNet [39]), and outperformed the training stage from scratch. DeepGaze II [40] has also been proposed based on a pre-trained model (VGG-19 [41]), where attention information was extracted from the VGGNet without fine-tuning the attention task. Next, the DeepFix model [42] was proposed by Kruthiventi et al. based on a pre-trained model VGG-16. Furthermore, in [43] object detection and saliency detection were carried out using a deep convolutional neural network (CNN). Finally, the SALICON net model [44] was proposed to capture multi-scale saliency using combined fine and coarse features from two-stream CNNs that were trained with multi-scale inputs.

Since the superior success of transfer learning models for visual saliency prediction has been established, several new models have been proposed that have improved saliency prediction performance. For instance, the SALICON model fine-tunes a mixture of deep features [44] using AlexNet [19], VGG-16 network [41], and GoogleNet [45] for visual saliency prediction. PDP [14] and DeepFix [42] were used on the VGG-19 network for the same task using MIT300 and the SALICON dataset, and FUCOS [46] fine-tunes features that were trained on the PASCAL dataset. Overall, DeepFix and SALICON models demonstrated significantly improved performance compared to DeepGaze I in the MIT benchmark.

4.3 Material and Methods

4.3.1 Proposed Model

The proposed model follows an FCN structure (i.e., a pixel-based approach) and the generic encoder-decoder form. The important difference between CNN and FCN networks is that the latter has learning filters throughout its structure. Even the decision-making layers at the end of the network are filters. FCNs also do not have any fully connected layers that are usually available at the end of the network.

Figure 4.1 explains the architecture of the proposed model for visual saliency prediction and the configuration of the proposed model is explained in Table 4.1. The encoder stage contains three blocks of convolution layers, each of which is followed by batch normalization, rectified linear unit (ReLU), and max pooling. The encoder stage is the same as that of a conventional CNN and generates feature maps by down-sample pooling. The decoder stage also transposes convolutional layers but does so in the opposite direction. Therefore, the decoder stage produces label maps (up-sampling) with the same input image size. The transposed convolution layers contain un-pooling and convolution operators. Unlike the max-pooling operation, the un-pooling operation increases the size of feature maps through the decoding stage. In addition, the image input size of the proposed model is 224 x 224 pixels.

Three inception modules are also used in the proposed model. Inception modules are useful because they benefit from different sized filters in one layer, which contributes to the multi-scale inference and enhances contextual information [20]. In addition, a residual module is

also added to the proposed model because it effectively avoids the vanishing gradient problem by introducing an identity shortcut connection [47] . Moreover, activations from a previous layer are reused by the residual module for the adjacent layer to learn its weights. Figure 2 shows the architecture of the inception and residual modules, respectively. Figure 4.2 (a) explains the layers of the inception module which contains three branches. The first two contain a sequence of two convolution filters, where the patch sizes of the layers are 1x1, the second layer is 3x3, and the last layer is 5x5, respectively. The third branch contains only one convolutional filter which has a patch size of 1x1. Each convolutional layer is followed by batch normalization and ReLU. Figure 4.2 (b) explains the structure of the residual module, which contains two branches. The first branch has a stack of three convolutional filters, sized 1x1, 3x3, and 1x1, respectively. The second branch has a single 1x1 convolutional filter. The two branches are combined by element-wise summation. Table 4.2 explains the number of each filter in the two modules (i.e., inception and residual). Notably, the convolutional module contains a Convolutional 2D, Batch Normalization, as well as a ReLU layer. The transposed convolutional module also contains the same layers as the convolutional module.

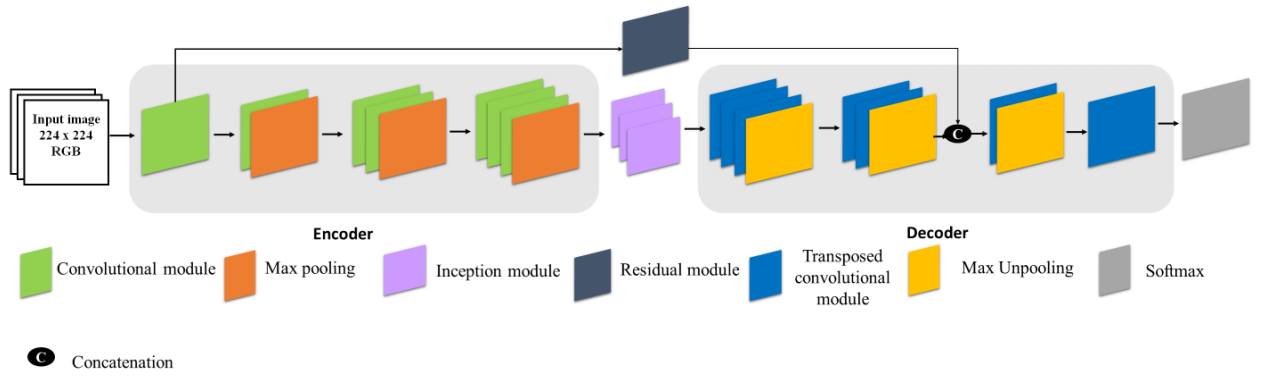


Figure 4.1. Architecture of the proposed model.

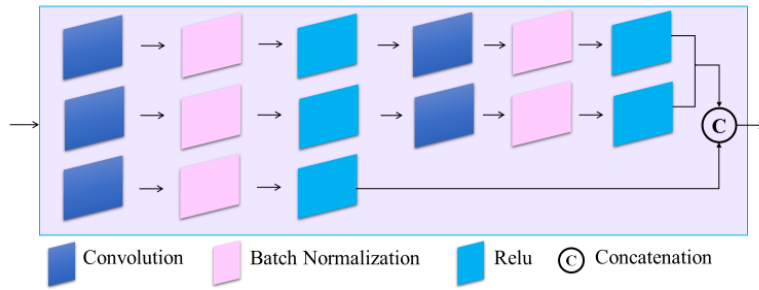
Table 4.1. Configuration of the proposed model.

-----	Layer type	Filter size
Encoder	Convolution	3x3, 64
	Residual Module	(*), 64
	Convolution	3x3, 128
	Max pooling	2x2
	Convolution	3x3, 256
	Max pooling	2x2
Decoder	Inception Module	(*), 256
	Transposed convolution	3x3, 256
	Convolution	3x3, 256
	Convolution	3x3, 128
	Transposed convolution	3x3, 64
	Convolution	3x3, 2
	Pixel Classification Layer	-

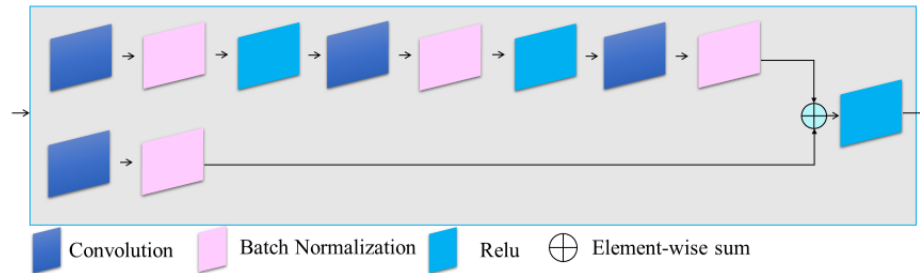
Note (*): See Table 4.2 for the filter size of the inception and residual modules.

Table 4.2. Configuration of inception and residual modules.

Module	Convolutional Configuration	Operation	Output
Inception	$1 \times 1, 256$ $1 \times 1, 256$ $3 \times 3, 256$ $5 \times 5, 256$ $1 \times 1, 256$	Concatenation	256
Residual	$1 \times 1, 32$ $3 \times 3, 64$ $1 \times 1, 64$	Element-wise sum	64



(a)



(b)

Figure 4.2. Architecture of (a) Inception and (b) residual modules.

4.3.2 Semantic Segmentation

The segmentation task plays an important role in image understanding and is essential for image analysis tasks. In semantic segmentation, each region or pixel is labeled as a class, such as flower, person, road, sky, ocean, or car. Many applications use semantic segmentation techniques, such as autonomous driving, Bio Medical Image Diagnosis, robotic navigation, localization, and scene understanding. Furthermore, Deep Neural Networks (DNNs) are commonly used as effective techniques for semantic segmentation [20]. Semantic segmentation works with semantics and location; global information determines the “what” while local information determines the “where” of an image. Deep feature hierarchies encode semantics and location in a nonlinear local-to-global pyramid [20]. Our proposed model (i.e., FCN) uses semantic segmentation techniques to assign each pixel in the given image into appropriate classes (i.e., foreground or background) in order to predict visual saliency (i.e., saliency map generation).

4.3.3 Datasets

The proposed model was trained using a standard available dataset (i.e., SALICON) and subsequently tested on four other well-established datasets, including TORONTO, MIT 300, MIT1003, and DUT-OMRON. All these datasets have different characteristics and so each is described below.

4.3.3.1 SALICON

The largest dataset for visual attention applications on the popular Microsoft Common Objects in Context (MS COCO) image database is SALICON [47]. This dataset contains 10,000 training, 5,000 validation, and 5,000 testing images with a fixed resolution of 480x640. While this dataset contains the ground truth data for the training and validation datasets, the test dataset ground truth data were unavailable [48].

4.3.3.2 TORONTO

One of the most widely used datasets for visual attention is the TORONTO dataset. It has 120 color images with a resolution of 511x681 pixels. This dataset contains images that were captured in indoor and outdoor environments and has been free-viewed by 20 human subjects [37].

4.3.3.3 MIT300

The MIT300 dataset has 300 natural images and the eye-tracking data of 39 users who free viewed these images were used to generate saliency maps. This dataset is challenging since its images are highly variable and natural [49]. A MIT saliency benchmark website for model evaluation (http://saliency.mit.edu/results_mit300.html) is available to evaluate any saliency model using this dataset.

4.3.3.4 MIT1003

MIT1003 includes 1,003 images from the Flickr and LabelMe collections. Saliency maps of these images have also been generated from the eye-tracking data of 15 users. This

dataset contains 779 landscape and 228 portrait images that vary in size from 405x405 to 1024x1024 pixels, making it the largest available eye fixation dataset [50].

4.3.3.5 DUT-OMRON

DUT-OMRON has 5,168 high quality images that were manually selected from over 140,000 images. The largest height or width of this dataset is 400 pixels and each image is represented by five subjects. There is more than one salient object in this type of dataset and the image has a more a complex background [51].

4.3.4 Evaluation Metrics

Several methods may be used to evaluate the correspondence between human eye fixation and model prediction [52]. Generally, saliency evaluation metrics are divided into distribution- and location-based metrics. Previous studies on saliency metrics found it is difficult to perform a reasonable comparison for assessing saliency models using a single metric [51]. Here, we accomplished our experiment by extensively considering several different metrics, including the Similarity Metric (SIM), Normalized Scanpath Saliency (NSS), and AUC. The last metric is the area under the receiver operating characteristic (ROC) curve (e.g., AUC-Borji, and AUC-Judd). For clarification, we indicate the map of fixation locations as Q , the predicted saliency map as S , and the continuous saliency map (distribution) as G .

4.3.4.1 Similarity Metric (SIM)

The SIM metric produces a histogram that is a measurement of the similarity between two distributions. This metric considers the normalized probability distributions of both the saliency and human eye fixation maps. SIM is also computed as the sum of the minimum values at each pixel, after normalizing the input maps. Equation (1) explains how to calculate the SIM metric.

$$SIM = \sum_{i=1} \min (\hat{S} (i), \hat{G} (i)), \quad (4.1)$$

where

$$\sum_i \hat{S} (i)=1, \text{ and } \sum_i \hat{G} (i)=1,$$

and \hat{S} and \hat{G} are the normalized saliency and the fixation maps, respectively. Importantly, a similarity of one indicates that the distributions are the same whereas a zero indicates that they do not overlap.

4.3.4.2 Normalized Scanpath Saliency (NSS)

NSS was is a simple correspondence measure between saliency maps and ground truth data, computed as the average normalized saliency at fixated locations. NSS is, however, susceptible to false positives and relative differences in saliency across the image [53]. To calculate NSS given a saliency map S and a binary map of fixation location F ,

$$NSS = \frac{1}{N} \sum_{i=1}^N \bar{S} (i) \times F(i), \quad (4.2)$$

where $N = \sum_i F(i)$ and $\bar{S} = \frac{S - \mu(s)}{\sigma(S)}$,

and N is the total number of human eye positions and $\sigma(S)$ is the standard deviation.

4.3.4.3 AUC-Borji

The AUC-Borji metric, based on Ali Borji's code [54], uses a uniform random sample of image pixels as negatives and defines false positives as any fixation (saliency) map values above the threshold of these pixels. The saliency map is a binary classifier that separates positive from negative samples at varying thresholds, the values of which are sampled at a fixed step size. The proportion of the saliency map values above the threshold at the fixation locations is the true positive (TP) rate. Conversely, the proportion of the saliency map values that occur above the threshold sampled from random pixels (as many samples as fixations, sampled uniformly from all image pixels) is the false positive rate (FP).

4.3.4.4 AUC-Judd

The AUC-Judd metric [50] is also popular for the evaluation of saliency models. As with AUC-Borji, positive and negative samples are separated at various thresholds by treating the saliency map as a binary classifier. Unlike AUC-Borji, however, the thresholds are sampled from the saliency map's values. The proportion of the saliency map's values above a specific threshold at specific fixation locations is known as the true positive (tp) rate. Alternatively, the proportion of the saliency map's values that occur above the threshold of non-fixated pixels is the false positive (fp) rate.

4.4 Experimental Results

This section explains all the steps for implementing our work (see Table 3 for more details about experimental steps). Specifically, training, adjusting the parameters, validating, and testing the proposed model on the aforementioned datasets (e.g., TORONTO, MIT300, MIT1003, and DUT-OMRON) are described in detail.

4.4.1 Model Training

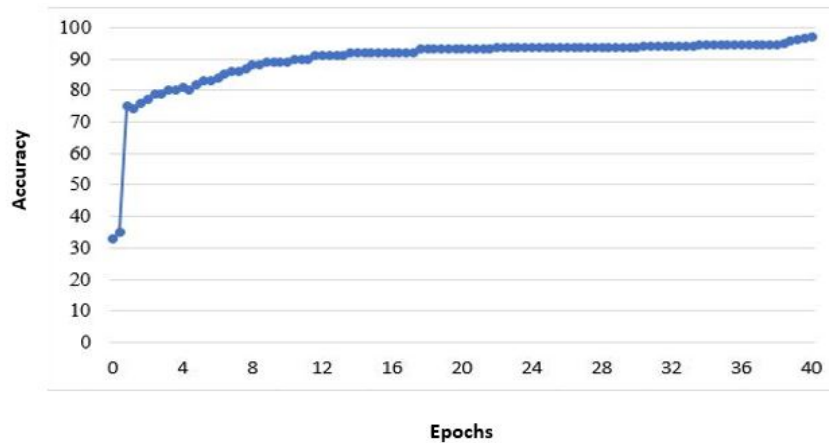
The most important step for the proposed deep learning model to predict visual saliency is model training. In this work, the proposed model was trained from scratch (i.e., full training). Training of the models from scratch is challenging due to computational and data availability, leading to problems of overfitting. However, there are several techniques, such as normalization, data augmentation, and dropout layers that are useful for mitigating the problems generated from overfitting.

In general, the full-training style has two different categories. In the first category, the CNN architecture is fully designed and trained from scratch. In this case, the number of CNN, pooling layers, the kind of activation function, neurons, learning rate, and the number of iterations should be determined. In the second category, the network architecture and the number of parameters remain unchanged, but the advantages of pre-existing architecture and full training is applied to given images.

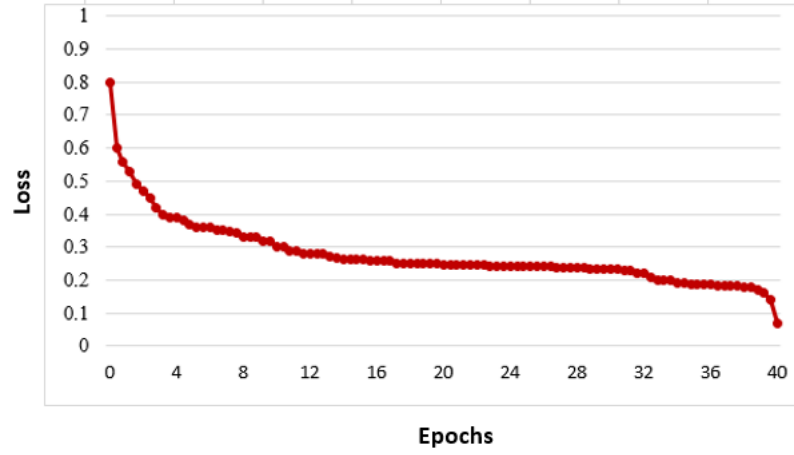
In this study, the first category was employed. Specifically, the proposed model was trained using the well-known dataset, SALICON (see Section 4.3.3.1 for more details) and was also validated using a specific validation dataset (i.e., 5000 images). This dataset is the

largest available for visual attention (i.e., 10,000 images for training, 5,000 for validation) and was created for saliency applications. At the beginning of the training task, all filter weights were randomly initialized because a pre-trained network was not used in this study. A mini batch of 16 images was used in each iteration and the learning rate was set as 0.001. The proposed model parameters were learned using the back-propagating loss function by stochastic gradient descent with a momentum (SGDM) optimizer

Since the number of images available for the training task was limited (i.e., 10,000 images), we suggested using the data augmentation technique to increase the number of training images by creating modified versions of images in the dataset. This technique was carried out to mitigate overfitting by rotating at 30° intervals. This technique also improves performance and the proposed model's ability for generalization. Figure 4.3 illustrates the proposed model's training progress from the mentioned training images (SALICON).



(a)



(b)

Figure 4.3. (a) Value of validation accuracy and (b) loss as a function of epochs.

4.4.2 Model Testing

In this step, we evaluated the proposed model using very well-known datasets including TORONTO, MIT300, MIT1003, and DUT-OMRON. Based on the experimental results, one can see the proposed model has the ability to predict visual saliency in a given image. The output of the test image is described as the saliency map, which can be obtained from the last layer of the proposed model. All the training and testing tasks were performed on an Intel CPU i7-3370K machine with 3.5 GHz and 16 GB RAM memory. An NVIDIA GeForce GTX 1080 Ti GPU with 11 GB of memory under CUDA version 8.0 was also utilized in this work.

4.5 Discussion

4.5.1 Quantitative Comparison of the Proposed Model with Other Advanced Models

To evaluate the efficiency of the proposed model for predicting visual saliency, we compared it to 10 state-of-the-art models, including ITT, AIM, Judd Model, GBVS, Mr-CNN, CAS, SalGAN, DeepGaze I, DeepGaze II, and ML-NET. The models were applied to the four previously mentioned datasets (i.e., TORONTO, MIT300, MIT1003, and DUT-OMRON), and the quantitative results are presented in Tables 4.3, 4.4, 4.5, and 4.6, respectively. All these models differ in terms of computational speed (i.e., run time). Table 4.7 explains the properties of the proposed model as well as the other 10 visual saliency models. From this table, one can see the run time of the proposed model is about 12 s on our machine (i.e., an Intel CPU i7-3370K).

Notably, the main difference between our proposed model and the other state-of-art models is that the proposed model was specifically designed for saliency prediction, whereas the other pre-trained models were essentially designed for object recognition and then fine-tuned for the visual saliency prediction task. In addition, our proposed model was trained from scratch, which requires a large number of training images to provide a reasonable performance; however, the largest dataset available for this application contains 10,000 images (e.g., SALICON) which was insufficient for training the proposed model from scratch.

Table 4.3 illustrates that, with the TORONTO dataset, the proposed model outperforms other models (deep and classical models) in terms of NSS; however, in terms of SIM, AUC-Judd, and AUC-Borji, the GBVS model provides the best results (note that the bolded values are the best results). From Table 4.4, one can see that with the MIT300 dataset, the model that provides the best performance is DeepGaz II in terms of the AUC-Judd and AUC-Borji metrics. However, the SalGAN model produces the best results for the SIM metric, while the ML-NET model provides the best value for the NSS metric. In Table 4.5 (for MIT1003 dataset), one can see that the proposed model surpasses the other models in terms of the SIM and AUC-Judd metrics, while the GBVS model provides the best results for the NSS metric. Finally, Table 4.6 shows that, with the DUT-OMRON dataset, the proposed model achieved the best result in terms of the AUC-Judd metric, while the GBVS model is the best in terms of the AUC-Borji metric.

Table 4.3. Comparison of the quantitative scores of several models on TORONTO [37] dataset.

Model	NSS	SIM	AUC-Judd	AUC-Borji
ITTI	1.30	0.45	0.80	0.80
AIM	0.84	0.36	0.76	0.75
Judd Model	1.15	0.40	0.78	0.77
GBVS	1.52	0.49	0.83	0.83
Mr-CNN	1.41	0.47	0.80	0.79
CAS	1.27	0.44	0.78	0.78
Proposed Model	1.52	0.46	0.80	0.76

Note. Humans baseline [36] 3.29 1.00 0.92 0.88

Table 4.4. Comparison of the quantitative scores of several models on MIT300 [49] dataset.

Model	NSS	SIM	AUC-Judd	AUC-Borji
ITTI [21]	0.97	0.44	0.75	0.74
AIM [55]	0.79	0.40	0.77	0.75
Judd Model [50]	1.18	0.42	0.81	0.80
GBVS [36]	1.24	0.48	0.81	0.80
Mr-CNN [26]	1.13	0.45	0.77	0.76
CAS [56]	0.95	0.43	0.74	0.73
SalGAN [57]	2.04	0.63	0.86	0.81
DeepGaze I [38]	1.22	0.39	0.84	0.83
DeepGaze II [58]	1.29	0.46	0.87	0.86
ML-NET [59]	2.05	0.59	0.85	0.75
Proposed Model	1.73	0.42	0.80	0.71

Table 4.5. Comparison of the quantitative scores of several models on MIT1003 [50] dataset.

Model	NSS	SIM	AUC-Judd	AUC-Borji
ITTI	1.10	0.32	0.77	0.76
AIM	0.82	0.27	0.79	0.76
Judd Model	1.18	0.42	0.81	0.80
GBVS	1.38	0.36	0.83	0.81
Mr-CNN	1.36	0.35	0.80	0.77
CAS	1.07	0.32	0.76	0.74
Proposed Model	1.35	0.44	0.88	0.78

Table 4.6. Comparison of the quantitative scores of several models on DUT-OMRON [57] dataset.



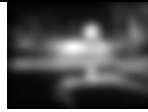


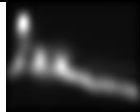














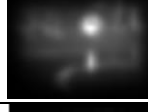

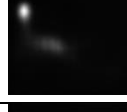
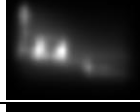





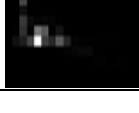
Model	NSS	SIM	AUC-Judd	AUC-Borji
ITTI	3.09	0.53	0.83	0.83
AIM	1.05	0.32	0.77	0.75
GBVS	1.71	0.43	0.87	0.85
CAS	1.47	0.37	0.80	0.79
Proposed Model	1.84	0.45	0.88	0.76

Table 4.7: Properties of the proposed model and ten visual saliency models

Model	Training	Deep Learning	Run Time
BMS	No	No	0.3 S
CAS	No	No	16 S
GBVS	No	No	2 S
ITTI	No	No	4 S
Mr-CNN	yes	Yes	14 S (GPU)
SalNet	yes	Yes	0.1 S (GPU)
eDn	yes	Yes	8 S (GPU)
AIM	yes	No	2 S
Judd Model	yes	No	10 S
DVA	yes	Yes	0.1 S (GPU)
Proposed Model	yes	Yes	12 S

4.5.2. Qualitative Comparison of the Proposed Model with Other Advanced Models

The qualitative results obtained by the proposed model are compared with those of five state-of-the-art models, such as ITTI, FES, CovSal, GBVS, and SDS-GM [60], on the aforementioned datasets (i.e., TORONTO, MIT300, MIT1003, and DUT-OMRON). Figure 4.4 explains the visual saliency map results and the proposed model predicts visual saliency, i.e., generate saliency map, within the given images. As shown in the figure, the proposed model has ability to consistently capture saliency from low-level features (e.g., colour) and more high-level features (e.g., human, face, and text). Based on the evaluation of the proposed model, we can see the proposed model produces reasonable saliency maps compared with other state-of-the-art models.

-----	TORONTO			MIT300		
Model	Test image	Ground Truth	Model prediction	Test image	Ground Truth	Model prediction
ITTI [7]						
FES [11]						
CovSal [12]						
GBVS [9]						
SDS-GM [13]						



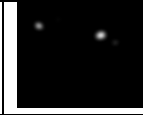

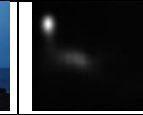
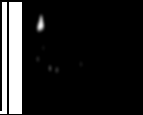

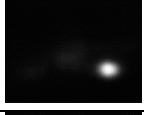
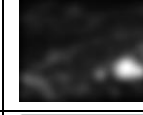

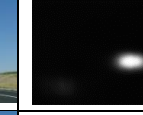
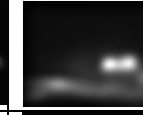

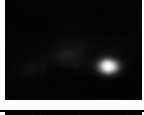
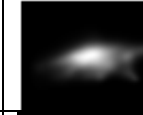

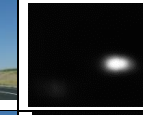
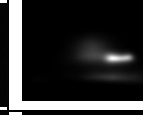

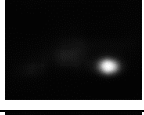
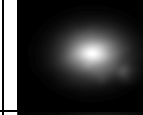

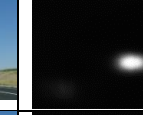
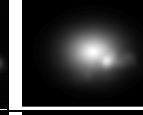

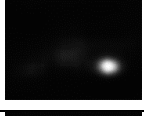
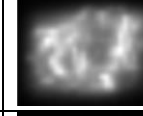

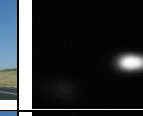


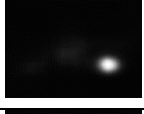
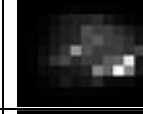

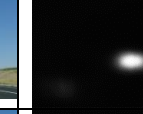


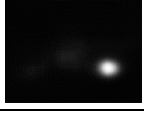
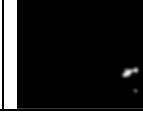

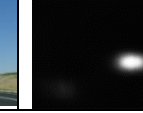
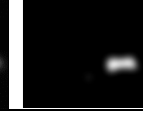
Proposed Model						
-----	MIT1003			DUT-OMRON		
ITTI						
FES						
CovSal						
GBVS						
SDS-GM						
Proposed Model						

Figure 4.4. Saliency maps obtained from the proposed model and five advanced models for a sample image from the TORONTO, MIT300, MIT1003, and DUT-OMRON datasets.

4.5.3 Ablation Study

In this work, we evaluated several different aspects of the proposed model's architecture. Table 4.8 illustrates the results of the experiments conducted in this work. Based on the architecture of the proposed model, we suggested 13 different scenarios in order to find an optimum architecture. Several conclusions were obtained based on these experiments:

- (1) From scenarios S1 to S4, we can see the best global accuracy is achieved with 3 encoder-3 decoder stages (i.e., global accuracy was 85.05 % and loss function was 0.2384).

(2) S7 describes the proposed model using 3 convolutional modules & 3 inception modules. This architecture also produced the best global accuracy (i.e., global accuracy was 93.63 %, and loss function was 0.1051) compared to S5 and S6, which contain one and two inception modules, respectively.

(3) S13 is the last scenario we selected as the entire model, including 3 convolutional, 3 inception, and 1 residual module (i.e., Figure 4.1). This scenario produced a higher global accuracy (i.e., global accuracy was 97.05 %, and loss function was 0.07) compare to those of scenarios S11 and S12.

Table 4.8. Different FCN models applied in this study.

FCN Models		Training		Validation	
Scenarios	Description	Accuracy	Loss	Accuracy	Loss
S1	2 convolutional modules	79.14 %	0.2650	78.88 %	0.2700
S2	3 convolutional modules	85.05 %	0.2384	83.08 %	0.2571
S3	4 convolutional modules	83.47 %	0.2548	82.94 %	0.2608
S4	5 convolutional modules	80.04 %	0.2873	76.52 %	0.2775
S5	3 convolutional modules & 1 inception modules	89.69 %	0.2119	85.05 %	0.2231
S6	3 convolutional modules & 2 inception modules	90.84 %	0.1995	85.37 %	0.2454
S7	3 convolutional modules & 3 inception modules	93.63 %	0.1051	89.24 %	0.1666
S8	3 convolutional modules & 1 residual modules	87.55 %	0.2138	84.97 %	0.2317
S9	3 convolutional modules & 2 residual modules	83.23 %	0.2597	82.10 %	0.2684

S10	3 convolutional modules & 3 residual modules	81.66 %	0.2750	79.12 %	0.2921
S11	3 convolutional modules & 1 inception module & 1 residual module	89.46 %	0.1829	88.59 %	0.1889
S12	3 convolutional modules & 2 inception module & 1 residual module	92.73 %	0.1255	89.92 %	0.2111
S13	3 convolutional modules & 3 inception module & 1 residual module	97.05 %	0.07	90.64 %	0.1588

4.6 Conclusions

A new deep CNN model has been proposed in this work for predicting visual saliency in the field of view. The main novelty of this model is its use of a new deep learning network with three encoders and three decoders (convolution and deconvolution) for visual saliency prediction, as well as its inclusion of two modules (inception and residual modules). The proposed model was trained from scratch and used the data augmentation technique to produce variations of images. The experiment results illustrate that the proposed model achieves superior performance relative to other state-of-the-art models. Moreover, we discovered that an increase in the number of training images will increase the model prediction accuracy (i.e., improvement in model performance); however, the implementation of the model requires a large amount of memory and so it is difficult to use large numbers of training images. Furthermore, because the model was trained from scratch, we expected the model will require more training data than other models, which are currently unavailable.

A promising direction for future research is to collect a new dataset, generate its ground truth, and design new models with good performance and improved evaluation metrics based on the one proposed herein. Extending the proposed model and applying it to

examples of dynamic saliency (i.e., video images), is another plausible and interesting avenue of research. The proposed model may also facilitate other tasks, such as scene classification, salient object detection, and object detection, making it applicable in a number of disciplines. Importantly, future models based on that proposed herein should be able to learn from high-level understanding, so they are able to, for example, detect the most important object of the image (e.g., focusing on the most important person in the room). Saliency models also need to understand high-level semantics in the visual scene (i.e., semantic gap), and cognitive attention studies can help to overcome some of the restrictions identified in the proposed model.

References

1. Liu, N. and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
2. Huang, J., et al., Integrating visual saliency and consistency for re-ranking image search results. IEEE Transactions on Multimedia, 2011. **13**(4): p. 653-661.
3. Lu, X. and X. Li, Multiresolution imaging. IEEE transactions on cybernetics, 2013. **44**(1): p. 149-160.
4. Cheng, G., et al., Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 2015. **53**(8): p. 4238-4249.
5. Lu, X., X. Li, and L. Mou, Semi-supervised multitask learning for scene recognition. IEEE transactions on cybernetics, 2014. **45**(9): p. 1967-1976.
6. Yao, X., et al., Semantic annotation of high-resolution satellite images via weakly supervised learning. IEEE Transactions on Geoscience and Remote Sensing, 2016. **54**(6): p. 3660-3671.
7. Gao, D. and N. Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. in Advances in neural information processing systems. 2005.

8. Mackenzie, A.K. and J.M. Harris, A link between attentional function, effective eye movements, and driving ability. *Journal of experimental psychology: human perception and performance*, 2017. **43**(2): p. 381.
9. Gao, D., V. Mahadevan, and N. Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. in *Advances in neural information processing systems*. 2008.
10. Le Meur, O., et al., A coherent computational approach to model bottom-up visual attention. *IEEE transactions on pattern analysis and machine intelligence*, 2006. **28**(5): p. 802-817.
11. Gao, D., S. Han, and N. Vasconcelos, Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009. **31**(6): p. 989-1005.
12. Kanan, C., et al., SUN: Top-down saliency using natural statistics. *Visual cognition*, 2009. **17**(6-7): p. 979-1003.
13. Fang, S., et al., Learning discriminative subspaces on random contrasts for image saliency analysis. *IEEE transactions on neural networks and learning systems*, 2016. **28**(5): p. 1095-1108.
14. Jetley, S., N. Murray, and E. Vig. End-to-end saliency mapping via probability distribution prediction. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

15. Kruthiventi, S.S., et al. Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
16. Pan, J., et al. Shallow and deep convolutional networks for saliency prediction. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
17. Vig, E., M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.
18. Girshick, R., et al. Rich feature hierarchies for accurate object detection and semantic segmentation. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
19. Krizhevsky, A., I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. in Advances in neural information processing systems. 2012.
20. Long, J., E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
21. Itti, L., C. Koch, and E. Niebur, A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence, 1998. **20**(11): p. 1254-1259.

22. Fu, K., et al., Normalized cut-based saliency detection by adaptive multi-level region merging. *IEEE Transactions on Image Processing*, 2015. **24**(12): p. 5671-5683.
23. Gong, C., et al. Saliency propagation from simple to difficult. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
24. Guo, F., et al., Video saliency detection using object proposals. *IEEE transactions on cybernetics*, 2017. **48**(11): p. 3159-3170.
25. Li, Y., et al. The secrets of salient object segmentation. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.
26. Liu, N., et al., Learning to predict eye fixations via multiresolution convolutional neural networks. *IEEE transactions on neural networks and learning systems*, 2016. **29**(2): p. 392-404.
27. Liu, Z., et al., Superpixel-based spatiotemporal saliency detection. *IEEE transactions on circuits and systems for video technology*, 2014. **24**(9): p. 1522-1540.
28. Liu, Z., W. Zou, and O. Le Meur, Saliency tree: A novel saliency detection framework. *IEEE Transactions on Image Processing*, 2014. **23**(5): p. 1937-1952.
29. Wang, W. and J. Shen, Deep visual attention prediction. *IEEE Transactions on Image Processing*, 2017. **27**(5): p. 2368-2378.

30. Wang, W., et al., Inferring salient objects from human fixations. IEEE transactions on pattern analysis and machine intelligence, 2019.
31. Wang, W., et al., Correspondence driven saliency transfer. IEEE Transactions on Image Processing, 2016. **25**(11): p. 5025-5034.
32. Wang, W., et al., Saliency-aware video object segmentation. IEEE transactions on pattern analysis and machine intelligence, 2017. **40**(1): p. 20-33.
33. Borji, A. and L. Itti, State-of-the-art in visual attention modeling. IEEE transactions on pattern analysis and machine intelligence, 2012. **35**(1): p. 185-207.
34. Wang, W., J. Shen, and L. Shao, Video salient object detection via fully convolutional networks. IEEE Transactions on Image Processing, 2017. **27**(1): p. 38-49.
35. Wang, W., et al., Revisiting video saliency prediction in the deep learning era. IEEE transactions on pattern analysis and machine intelligence, 2019.
36. Harel, J., C. Koch, and P. Perona. Graph-based visual saliency. in Advances in neural information processing systems. 2007.
37. Bruce, N. and J. Tsotsos. Saliency based on information maximization. in Advances in neural information processing systems. 2006.
38. Kümmerer, M., L. Theis, and M. Bethge, Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. arXiv preprint arXiv:1411.1045, 2014.

39. Deng, J., et al. Imagenet: A large-scale hierarchical image database. in 2009 IEEE conference on computer vision and pattern recognition. 2009. Ieee.
40. Kümmerer, M., T.S. Wallis, and M. Bethge, DeepGaze II: Reading fixations from deep features trained on object recognition. arXiv preprint arXiv:1610.01563, 2016.
41. Simonyan, K. and A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
42. Kruthiventi, S.S., K. Ayush, and R.V. Babu, Deepfix: A fully convolutional neural network for predicting human eye fixations. IEEE Transactions on Image Processing, 2017. **26**(9): p. 4446-4456.
43. Mahadevan, V. and N. Vasconcelos. Saliency-based discriminant tracking. in 2009 IEEE conference on computer vision and pattern recognition. 2009. IEEE.
44. Huang, X., et al. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. in Proceedings of the IEEE International Conference on Computer Vision. 2015.
45. Szegedy, C., et al. Going deeper with convolutions. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
46. Bruce, N.D., C. Catton, and S. Janjic. A deeper look at saliency: Feature contrast, semantics, and beyond. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

47. Lin, T.-Y., et al. Microsoft coco: Common objects in context. in European conference on computer vision. 2014. Springer.
48. Jiang, M., et al. Salicon: Saliency in context. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
49. Judd, T., F. Durand, and A. Torralba, A benchmark of computational models of saliency to predict human fixations. 2012.
50. Judd, T., et al. Learning to predict where humans look. in 2009 IEEE 12th international conference on computer vision. 2009. IEEE.
51. Riche, N., et al. Saliency and human fixations: State-of-the-art and study of comparison metrics. in Proceedings of the IEEE international conference on computer vision. 2013.
52. Ghariba, B., M.S. Shehata, and P. McGuire, Visual Saliency Prediction Based on Deep Learning. *Information*, 2019. **10**(8): p. 257.
53. Bylinskii, Z., et al., What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 2018. **41**(3): p. 740-757.
54. Borji, A., et al. Analysis of scores, datasets, and models in visual saliency prediction. in Proceedings of the IEEE international conference on computer vision. 2013.

- 55. Bruce, N.D. and J.K. Tsotsos, Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, 2009. **9**(3): p. 5-5.
- 56. Goferman, S., L. Zelnik-Manor, and A. Tal, Context-aware saliency detection. *IEEE transactions on pattern analysis and machine intelligence*, 2011. **34**(10): p. 1915-1926.
- 57. Pan, J., et al., Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.
- 58. Kummerer, M., et al. Understanding low-and high-level contributions to fixation prediction. in *Proceedings of the IEEE International Conference on Computer Vision*. 2017.
- 59. Cornia, M., et al. A deep multi-level network for saliency prediction. in *2016 23rd International Conference on Pattern Recognition (ICPR)*. 2016. IEEE.
- 60. Li, Y. and X. Mou, Saliency detection based on structural dissimilarity induced by image quality assessment model. *Journal of Electronic Imaging*, 2019. **28**(2): p. 023025.
- 61. Borji, A., Saliency prediction in the deep learning era: An empirical investigation. *arXiv preprint arXiv:1810.03716*, 2018.
- 62. Borji, A., H.R. Tavakoli, and Z. Bylinskii, Bottom-up Attention, Models of. *arXiv preprint arXiv:1810.05680*, 2018.

Chapter 5. Salient Object Detection Using Semantic Segmentation

Techniques

Abstract

Salient Object Detection (SOD) is the operation of detecting and segmenting a salient object in a natural scene. Several studies have examined various state-of-the-art machine learning approaches for SOD. In particular, Deep Convolutional Neural Networks (CNNs) are commonly applied for SOD because of their powerful feature extraction abilities. In this work, we investigate the semantic segmentation capability of several well-known pre-trained models, including FCNs, VGGs, ResNets, MobileNet-v2, Xception, and InceptionResNet-v2. These models have been trained over an ImageNet dataset, fine-tuned on a MSRA-10K dataset, and evaluated using other public datasets, such as ECSSD, MSRA-B, DUTS, and THUR15k. The results illustrate the superiority of ResNet50 and ResNet18, which have Mean Absolute Errors (MAE) of approximately 0.93 and 0.92, respectively, compared to other well-known FCN models. Moreover, the most robust model against noise is ResNet50, whereas VGG-16 is the most sensitive, relative to other state-of-the-art models.

5.1 Introduction

The Human Visual System (HVS) has the ability to detect visually distinguished stimuli (the pre-attentive stage) called salient regions. The filtered salient regions are then processed in more detail to obtain high-level information (the attentive stage). The

detection of salient objects has long been studied by neuroscientists and psychologist and has recently become of interest to the computer vision community. The task of Salient Object Detection (SOD) is very useful, and works as a pre-processing step for image processing and computer vision tasks, such as object detection [1], image classification [2], video summarization [3], content-aware image editing [4], image understanding [5, 6], and image captioning [7, 8].

In the computer vision field, the SOD model detects the most salient object in the scene and segments the accurate region of that object. In other words, the SOD model aims to localize and extract the most prominent and eye-attracting objects or regions in an image. Most of the existing models attempt to extract the most salient object; however, some of these models can be used to find many salient objects in an image. Generally, the SOD model performs well if it can meet the following criteria: 1) the probability of the real salient object region and the reality of marking the foreground as a salient object should be high; 2) high computational efficiency; the SOD model should quickly detect salient regions; and 3) high resolution; the output of the SOD model should have full resolution and maintain the information of the original image [9].

Deep Convolutional Neural Networks (CNNs) have been recently adopted for SOD because they can extract robust features, achieving results comparable to other state-of-the-art methods [6]. In addition, a feature extraction process based on CNNs includes more high-level semantic information, since the CNNs are pre-trained on the datasets object recognition tasks. In particular, Fully Convolutional Neural networks (FCNs) have recently

been utilized for SOD. This type of deep network was proposed by Long et al. [10] and has produced impressive results for SOD because it has the ability to address pixel-based classification tasks in an end-to-end style. Furthermore, FCN has the same architecture as CNN, but does not contain any fully connected layers. FCNs are also powerful visual models that produce hierarchies of features by generating high-level features from low-level features. Since Long et al propose the original process of semantic segmentation based on FCN, three novel architectures, including FCN-8s, FCN-16s, and FCN-32s, have been developed.

Despite deep improvements having been made, there are still two major challenges that prevent its use in real-world application, such as embedded devices. The first challenge is that FCN-based models produces the low resolution of the saliency map. Due to the pooling operations and repeated stride in CNN, it is unavoidable to lose resolution and complicated to smooth. Therefore, it is impossible to locate salient objects precisely, particularly for the small objects and object boundaries. The second challenge is the dense weight and large redundancy of current deep saliency models. Consequently, this task is too heavy for the pre-processing step to apply in the subsequent high-level tasks; and also, the memory is not sufficient for embedded devices.

This work evaluates the capabilities and accuracies (i.e., transfer learning, fine tuning) of ten pre-trained models based on FCN (i.e., semantic segmentation technique) to address the problem of SOD. These models include VGG-Net (VGG-16, VGG-19), (FCN-8S, FCN-16S, FCN-32S), ResNet-18, ResNet-50, MobileNet-v2, Xception, and

InceptionResnet-v2. Most of these models were trained for image classification tasks on more than a million images for from the ImageNet database [11].

In summary, this work has the following contributions:

1. We investigate FCNs (i.e., semantic segmentation) for SOD.
2. All pre-trained models (i.e., FCNs, VGGs, ResNets, MobileNet-v2, Xception, and InceptionResNet-v2) were evaluated on four well-known benchmark datasets, including ECSSD, MSRA-B, DUTS, and THUR15k.
3. Sensitivity analysis was performed to evaluate the trained models against noise.

The remainder of this chapter is organized as follows: Section 5.2 explains the related work of this research. Section 5.3 explains the materials and methods used in this work. Section 5.4 describes the experimental results. Section 5.5 presents the quantitative and qualitative experimental results, and the assessment of the trained models against noise over four benchmark datasets. Finally, our conclusions are drawn in Section 5.6.

5.2 Related Work

Salient Object Detection has received attention from computer vision researchers for many years. Early methods try to focus on low-level features and cues. The most widely applied is Contrasted Prior, which considers that salient regions make high contrast compared with the background in the scene [12-17]. Additionally, there is an approach that uses color uniqueness and spatial distribution to compute saliency of objects [18]. Jiang et al. [19] and Zhang et al. [20] use perspective of objects' uniqueness and

surroundings to detect salient objects. Shen et al. [21] consider the background can be demonstrated by a low-rank matrix, and salient objects are the sparse noise. In addition, an approach called (Center Bias) that supposes the salient object is located in the center of the image [20, 22-25].

In the last few years, many SOD models have been proposed, especially with CNNs employed in this task. Recently, CNNs have shown superior performance in many tasks of computer vision because they have ability to extract high-level and multi-scale features. Li et al. [26] suggested the use of multi-scale features extracted from a deep CNN to extract a saliency map. Wang et al. [27] use integrating both local estimation and global search for predicting a saliency map; two different deep CNNs are trained to capture global contrast and local information. Zhao et al. [28] proposed multi-context deep learning framework for SOD. They applied two CNNs to extract local and global context information. Lee *et al.* [29] considered both hand-crafted features and high-level features. To merge these features together, a combined fully connected neural network was designed in order to generate the saliency map. Liu and Han [30] designed a two-stage deep network. The first stage produced a coarse prediction map, and the second stage refine the details the output of the first stage (i.e., the coarse prediction map) hierarchically and gradually. Li and Yu [31] designed a deep contrast network. This network merged a segment-wise spatial pooling stream and a pixel-level fully convolutional stream.

FCN-based methods are another approach for SOD [31, 32]. These achieve an important improvement compared with patch-wise deep networks methods, because FCN has the ability to capture richer and multiscale information. Hu *et al.* [33] proposed to learn a level set [34] function to output accurate boundaries and compact saliency. Luo *et al.* designed a network with a 4 x 5 grid structure to combine global and local information, then used a fusing loss of cross entropy and boundary IoU inspired by [35]. Zhang *et al.* [36] proposed a sibling architecture and a structural loss function for predicting saliency with clear boundaries. Zhang *et al.* [37] designed a controlled bi-directional passing of features between layers and shallow layers to achieve accurate predictions. In spite of tremendous efforts and huge progress in the last two years, there is still ample room for improvement over the generic CNN models for SOD.

5.3 Materials and Methods

5.3.1 Semantic Segmentation Techniques

In general, image segmentation is the process of partitioning a digital image into multiple portions (i.e., sets of pixels, also known as image objects). Semantic segmentation, in particular, is a novel technique in the field of computer vision. The task of image segmentation is useful because it simplifies and changes the representation of an image to something easier to analyze [38]. Looking at the larger picture, semantic segmentation is a high-level task that smooths the way towards a complete scene understanding. Semantic segmentation can be defined as classification of the object class for each pixel within an image [39]. That means there is a label (e.g., flower, person, road, sky, ocean, or car) for

each pixel. In other words, semantic segmentation uses a pixel classification layer to predict the categorical label for every pixel in an input image. We can, therefore, think of semantic segmentation as image classification at the pixel level. Many applications now utilize semantic segmentation techniques, such as autonomous driving [40], industrial inspection [41], classification of terrain visible in satellite imagery [42], and medical imaging analysis [43]. The next section describes the pre-trained models that have been used to support the semantic segmentation technique.

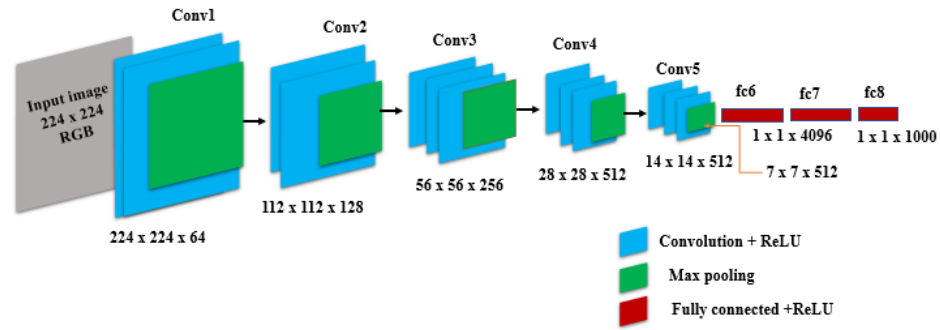
5.3.2 Pre-Trained Models

Using a pre-trained model for fine-tuning is a good solution when only a small dataset is available. All pre-trained models used in this study have been trained on a huge dataset (e.g., the ImageNet dataset [11]) and can classify images into 1000 object classes, such as a keyboard, mouse, pencil, and many animals. Consequently, the networks have learned rich feature representations for a wide range of images, subsequently adjusting the parameters for the SOD dataset. In this subsection, we briefly review ten state-of-the-art pre-trained models that we use later in this work.

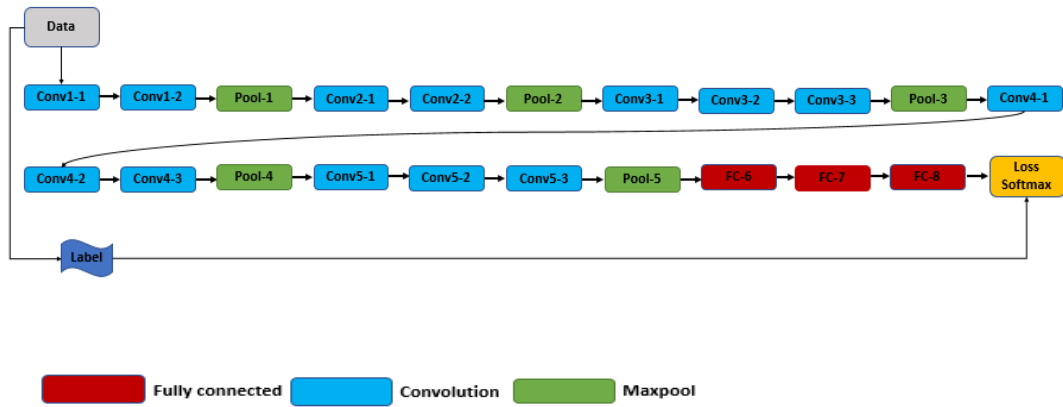
1.VGG-16 Network

The VGG-16 network was developed by Simonyan and Zisserman in the 2014 ImageNet large Scale Visual Recognition Competition (ILSVRC) competition [44]. Generally, the VGG-16 network contains 13 convolution layers, five pooling layers, and three fully connected layers. The VGG-16 network can classify images into 1000 object classes and

has an image input size of 224×224 . Figure 5.1 (a, b) illustrates the general structure and the data flow through the VGG-16 network.



(a)



(b)

Figure 5.1: General Structure of VGG-16 network: (a) Convolution layers and (b) data flow [14].

The major difference between the VGG-16 network and previous networks is the use of a series of convolution layers with small receptive fields (3×3) in the first layers instead of a few layers. This results in fewer parameters and more nonlinearities in between, making the decision function more selective and the model easier to apply for training [44].

The input image is passed over a series of convolution layers with 3×3 convolutional filters. This is beneficial because the filter will capture the notation of the center, left/right, and up/down. The convolution stride is set to 1 pixel, whereas the padding is set to 1 pixel. Five max-pooling layers are used after convolution layers for the down-sampling operation (i.e., dimensionality reduction). Each max-pooling is also performed over 2×2 pixels, with a stride value of 2. In addition, three fully connected (FC) layers follow a series of convolution layers. Specifically, the first two have 4096 channels each, and the third has 1000 channels. The structure of the fully connected layers is the same in all networks. The final layer is a soft-max layer that must have the same number of nodes as the output layer. The function of the soft-max layer is to map the non-normalized output to a probability distribution through predicted output classes [44], [45].

2.VGG-19 Network

VGG-19 is a CNN that contains 16 convolution layers, five pooling layers, and three fully connected layers. This network has 19 layers and has an image input size of 224×224 [44]. This pre-trained network is trained on more than a million images from the ImageNet database [11]. This network has ability to classify images into 1000 classes, such as mouse, keyboard, pencil, and animals. In addition, this increases the depth of the network and

contributes to learning more complex features. The impressive results of VGG revealed that the network depth is a significant factor in obtaining high classification accuracy.

Figure 5.2 explains the data flow in the VGG-19 network.

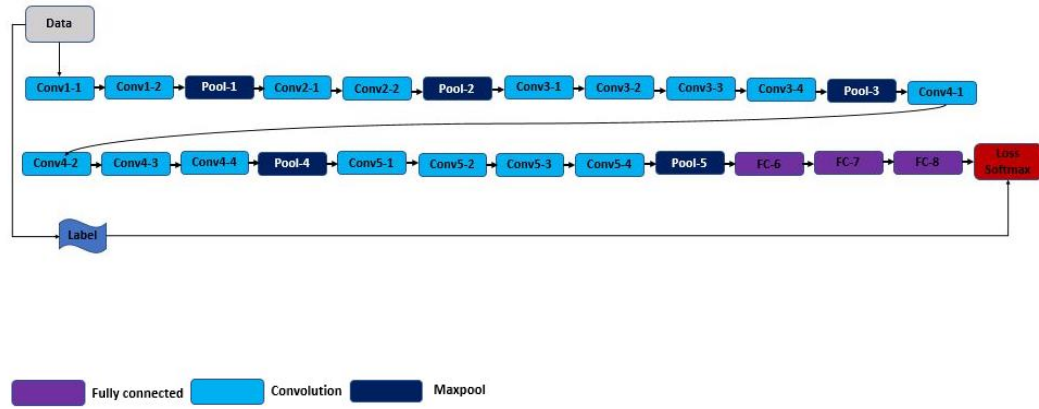


Figure 5.2: Data flow in the VGG-19 Network.

3. ResNet-18 Network

ResNet-18 is a CNN developed by Kaimming et al.; it won first place in the ILSVRC2015 competition [46]. This network has 18 layers and has an image input size of 224 x 224. [11]. This network can classify images into 1000 classes, such as mouse, keyboard, pencil, and animals. Table 5.1 explains the architecture of the ResNet-18 network [47].

Table 5.1: ResNet-18 Architecture.

Layer Name	Output Size	ResNet-18
Conv1	112 x 112 x 64	7 x 7, 64, stride 2
Conv2_x	56 x 56 x 64	3 x 3 max pool, stride 2 $\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$
Conv3_x	28 x 28 x 128	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$
Conv4_x	14 x 14 x 256	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$
Conv5_x	7 x 7 x 512	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$
average pool	1 x 1 x 512	7 x 7 average pool
fully connected	1000	512 x 1000 fully connection
Softmax	1000	-----

4. ResNet-50 Network

ResNet-50 is a CNN with 50 layers and, as with the ResNet-18 and VGG-19 networks, can classify images into 1000 classes, such as a mouse, pencil, keyboard, and animals [46]. As a result, the network has learned rich feature representations for a wide range of images. The network has image input size of 224 x 224. Figure 5.3 explains the architecture of this network.

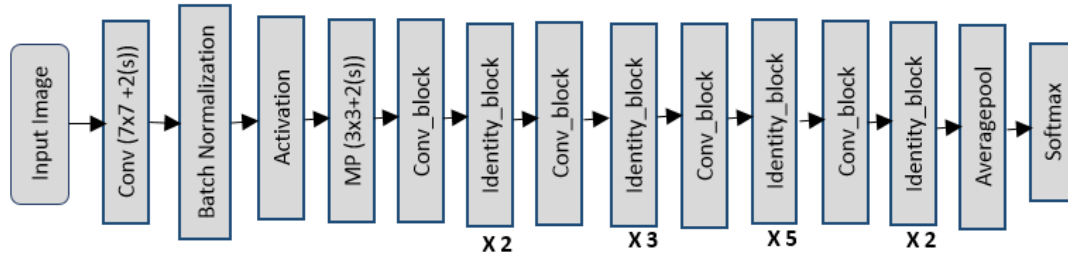


Figure 5.3: ResNet-50 Network Architecture [17].

5. MobileNet-v2 Network

MobileNet-v2 Network is a CNN with 54 layers and has an image input size of 224×224 . In addition, this network is the second version (V2) of MobileNet, and so is more efficient and powerful than the original. This version contains two types of blocks: (1) a residual block with a stride of 1 and (2) a block with a stride of 2 for downsizing [49]. Figure 5.4 illustrates the main building block of MobileNet-v2 network.

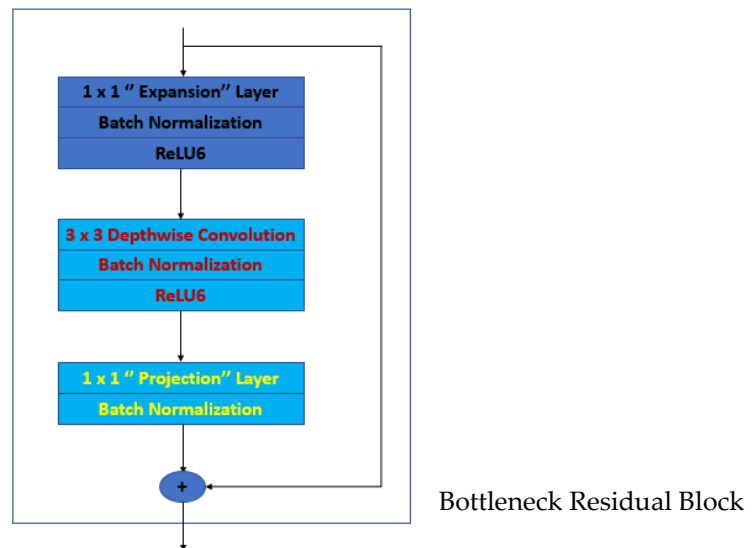


Figure 5.4: Main block of MobileNet-v2 Network.

6. Xception Network

Xception is a CNN presented by Francois Chollet at the Conference on Computer Vision and Pattern Recognition (CVPR) in 2017. This network has 18 layers and an image input size of 299×299 . Figure 5.5 shows the architecture of Xception network [50]. This network

has an architecture, constructed based on a linear stack of a depth-wise separable convolution layer with linear residual connections. In this configuration, there are two significant convolution layers: a depth-wise convolution layer [51], where a spatial convolution is executed independently in each channel of input data, and point-wise convolutional layer, where a 1×1 convolutional layer maps the output channels to a new channel space using a depth-wise convolution.

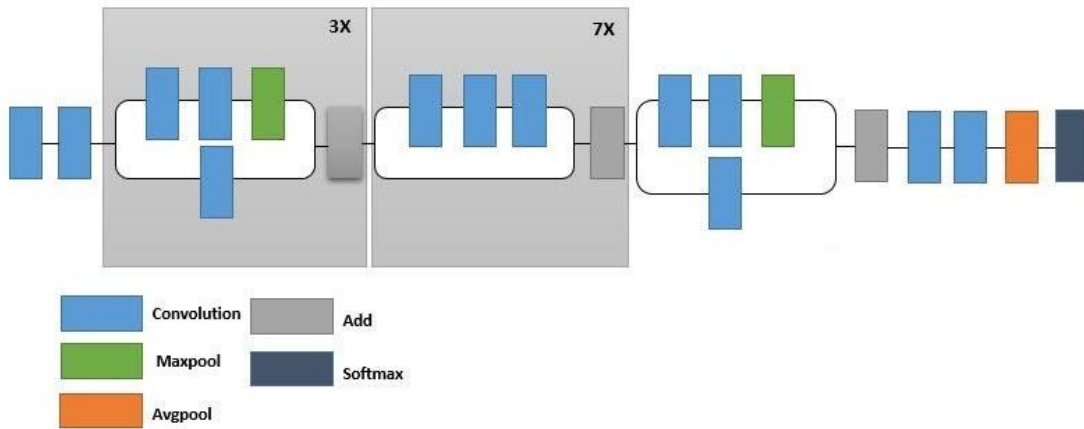


Figure 5.5: The architecture of Xception Network.

7. InceptionResnet-v2 Network

InceptionResNet-v2 is a CNN designed based on two successful deep networks, ResNet [46] and Inception [52]. This network has 164 layers and an image input size of 229×229 [53]. Batch-normalization is used only on the top of the traditional layers, rather than on the top of the summation. In particular, residual modules are used to allow an increase in

the number of Inception blocks, therefore, increasing the network depth. Moreover, the most important problem associated with deep networks is the training stage. This problem can be fixed using residual connections [46]. The residual connection is an efficient approach to solving the training problem, especially when a large number of filters is applied in the network. Subsequently, scaling the residual connections contributes to stabilizing the network during the training stage. Figure 5.6 explains the architecture of the InceptionResnet-v2 used in this work.

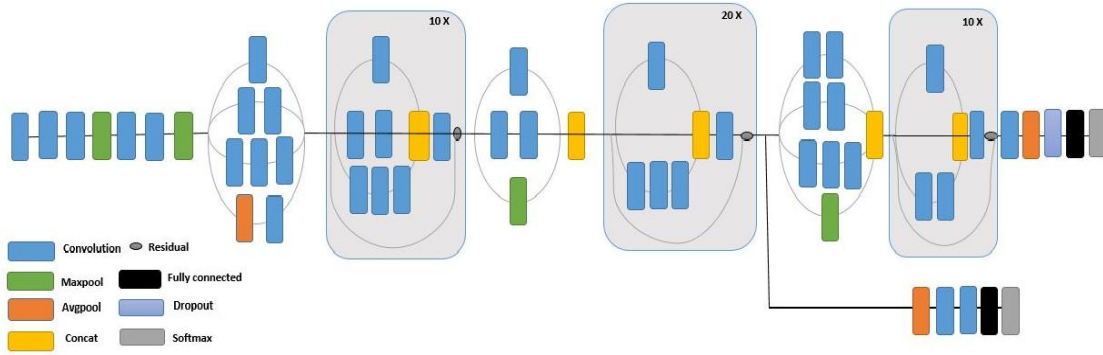


Figure 5.6: The architecture of InceptionResnet-v2 Network.

8. Fully Convolutional Networks (FCN)

A FCN is a neural network containing convolutional layers, none of which are fully connected and are always available at the end of the network. Long, J et al., 2015 were the first to use a FCN to transform image pixels to pixel categories [10]. The main difference

between the classical CNN and FCN is that the latter has learning filters throughout its structure; even the decision-making layer at the end of the FCN contains filters.

There are three different forms of FCN architecture that differ in the spatial precision of their output: FCN-8s, FCN-16s, and FCN-32s. Figure 5.7 explains the architecture of the FCN-8s, FCN-16s, and FCN-32s.

1. **FCN-8S:** This type sums the 2X up-sampled conv7 with pool 4, up-samples them with stride 2 transposed convolution, and sums them with pool 3, prior to applying transposed convolution with stride 8 to produce the segmentation map.
2. **FCN-16s:** This type sums the 2X up-sampled prediction from conv 7 with pool 4 to generate the segmentation map. This uses a transposed convolution layer with stride 16.
3. **FCN-32s:** This type generates the segmentation based on the output of conv7 and uses a transposed layer with stride 32.

Figure 5.7 explains the architecture of the FCN-8s, FCN-16s, and FCN-32s models.

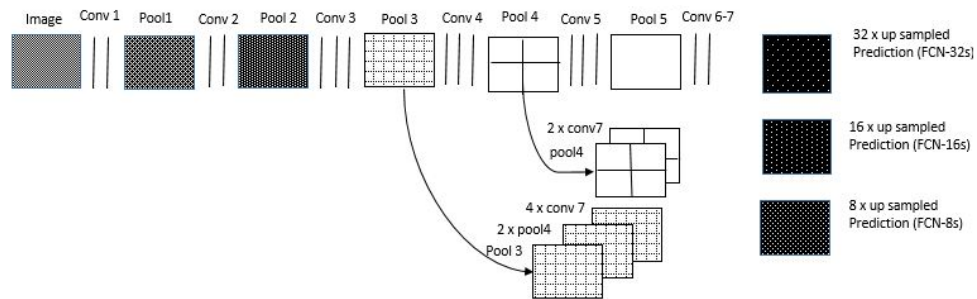


Figure 5.7: FCN models architecture [10].

5.3.3 Datasets

Since the development of SOD models, many datasets have been presented. These datasets play an important role in model training and evaluating performance (i.e., testing). In this work, we selected the dataset MSRA-10K for model training and five datasets (ECSSD, MSRA-B, MSRA-10K, DUTS, and THUR15k) for model testing.

5.3.3.1 ECSSD

This dataset contains 1000 images (JPG format). It has a semantic meaning and complex natural contents [54].

5.3.3.2 MSRA-B

This dataset is a subset of MSRA-A and contains 5000 images. The images were relabeled by nine users using bounding boxes [16].

5.3.3.3 MSRA-10K

This dataset is known as THUS10K. It contains 10000 images [18] selected from MSRA [16] and covers all 1000 images in ASD [55]. All the images have a fixed bounding box. This dataset has been widely used for SOD models because it has a large scale and accurate annotation.

5.3.3.4 DUTS

This dataset is the largest used for SOD models. It contains 10553 images for training (JPG format) and 5019 images for testing (JPG format). The training images have been selected from ImageNet train/val set [56] and the test images from the ImageNet set [56].

5.3.3.5 THUR15k

This dataset contains about 15000 images (JPG format) [30]. It is introduced to evaluate sketch-based image retravel. This dataset has also been divided into five subset datasets, including a butterfly, coffeeMug, dogjump, giraffe, and plane.

5.3.4 Evaluation Metrics

This dataset contains about 15000 images (JPG format) [30]. It is introduced to evaluate sketch-based image retravel. This dataset has also been divided into five subset datasets, including a butterfly, coffeeMug, dogjump, giraffe, and plane.

5.3.4.1 Precision-Recall (PR)

A PR metric is computed by comparing the salient map (binary mask) and ground truth with the following equation:

$$precision = \frac{Tp}{Tp+FP}, Recall = \frac{TP}{TP+FN}, \quad (5.1)$$

where Tp , TN , FP , and FN are true-positive, true-negative, false positive, and false negative, respectively.

A set of thresholds that range from 0 to 255 are used to generate a binary mask, thus producing a pair of precision/recall values to evaluate model performance.

5.3.4.2 F-measure

This metric is computed based on precision and recall by calculating a weighted harmonic mean as illustrated in the following equation:

$$F_B = \frac{(1+B^2)precision \times Recall}{B^2 precision + Recall}, \quad (5.2)$$

where B^2 is selected as 0.3 empirically to give more weight to precision [55]. Recall rate is also not as important as precision.

5.3.4.3 Mean Absolute Error (MAE)

This metric uses average pixel-wise absolute error between the ground truth $G \in \{0,1\}^{W \times H}$ and normalized map $S \in [0,1]^{W \times H}$. However, these fail to take into consideration the true negative pixels. Equation 3 explains how to calculate the MAE metric [17].

$$MAE = \frac{1}{W \times H} \sum_{i=1}^W \sum_{j=1}^H |G(i,j) - S(i,j)|, \quad (5.3)$$

where the W and H is the width and height of the given image.

5.3.4.4 Weighted F_B measure (Fbw)

This metric is an extension of the F-measure metric [28]. It attempts to modify the four parameters, TP, TN, FP, and FN, to real values and applies various weights (w) to different errors in numerous locations using the following equation:

$$F_B^w = \frac{(1+B^2)precision^w \times Recall^w}{B^2 precision^w + Recall^w}. \quad (5.4)$$

5.4 Experimental Results

In this study, semantic segmentation techniques were applied to ten pre-trained models, including VGG-Net (VGG-16, VGG-19), Fully Convolutional Network (FCN), (FCN-8S, FCN-16S, FCN-32S), ResNet-18, ResNet-50, MobileNet-v2, Xception, and InceptionResnet-v2. This section describes the training and model testing results.

5.4.1 Model Training

Since this work is based on pre-trained models, a large training dataset is not required. Here, all pre-trained models were trained over a well-known dataset, MSRA-10K (see Section 5.3.3. for more details). In the training stage, 1000 images were employed. Table 5.2 explains the important parameters (e.g., training global accuracy, validation accuracy, training time, etc.) provided after training. The investigated models were trained using minimization of the cross-entropy loss function and stochastic gradient descent with a

momentum (SGDM) optimizer. The mini-batch size was 10, the number of epochs was 25, and the learning rate was set at 0.001.

Table 5.2: The most important parameters through the training stage of all pre-trained models.

-----	Training		Validation		Training Time	Testing Accuracy (%)
Pre-Trained Model	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss	min, Sec	-----
VGG-16	96.95 %	0.1209	91.07 %	0.2345	1515, 37	91.21
VGG-19	96.95 %	0.1187	91.03 %	0.2377	1780, 27	91.34
(ResNet-18)	99.14 %	0.0212	93.35 %	0.2991	2425, 19	93.44
(ResNet-50)	99.37 %	0.0155	94.70 %	0.2602	1550, 48	95.37
MobileNet-v2	98.75 %	0.0332	91.83 %	0.3487	1102, 01	92.67
Xception	99.20 %	0.0202	92.55 %	0.3572	2250, 55	92.03
InceptionResnet-v2	99.60 %	0.0100	93.25 %	0.3499	2621, 59	92.97
FCN-8S	94.94 %	0.1351	90.14 %	0.2559	1526, 40	90.14
FCN-16s	95.90 %	0.0949	92.87 %	0.1960	1467, 58	92.87
FCN-32s	96.76 %	0.0732	92.06 %	0.2228	4482, 45	92.06

Note: Bolded values represent the most efficient model for each parameter.

5.4.2 Model Testing

During the testing stage, we selected sample datasets to evaluate all trained models. Once given a query image, the salient object was detected from the last layer of the deep model.

The processes of training and testing were carried out on an Intel CPU i7-3770k machine

with 3.5 GHz and 16 GB RAM. Table 5.3 explains the most relevant parameters (i.e., depth, image input size, number of parameters, and segmentation time) influencing the training and testing time. Note that the segmentation time is different for all trained models.

Table 5.3: Several parameters that influence the training and testing of different models in this study. **Note:** M means Million.

Pre-Trained Models	Depth	Image Input Size	Number of Parameters (M)	Segmentation Time (Sec)
VGG-16	16	224 -by- 224	138	5.16
VGG-19	19	224 -by- 224	144	5.43
ResNet-18	18	224 -by- 224	11.7	5.51
ResNet-50	50	224 -by- 224	25.6	7.50
MobileNet-v2	53	224 -by- 224	3.5	4.72
Xception	71	299 -by- 299	22.9	6.28
InceptionResnet-v2	164	299 -by- 299	55.9	24.78
FCN-8S	18	224 -by- 224	2.06	8.24
FCN-16s	17	224 -by- 224	--	7.69
FCN-32s	16	224 -by- 224	--	7.56

5.5 Discussion

5.5.1 Quantitative Comparison of the State-of-the-Art Pre-Trained Models for SOD

To recognize the best pre-trained model for SOD, we compared 10 that are commonly employed for semantic segmentation. All models were trained on a MSRA10K dataset and

were evaluated on four well-known datasets (ECSSD, MSRA-B, DUTS, and THUR15k). Tables 5.4 and 5.5 illustrate the evaluation metrics of all pre-trained models applied to the previously mentioned datasets. Based on the obtained results, ResNet-50 performed the best and FCN-8S performed the worst compared to other models. In particular, FCN-32s, ResNet-50, and InceptionResnet-v2 presented in Table 5.4 (for ECSSD dataset) showed a better performance compared to other models of the precision metric (pre). In addition, FCN-8s, FCN-16s, and Xception presented the worst performance compared to other models on the same metric. In the same table (Table 5.4), we can see the Recall metric (Recall) showed that the ResNet-50, MobileNet-v2, and Xception are the best models compared to other models. Also, the VGG-19, FCN-32s, and VGG-16 presented the weakest performance compared to other models. For the measure (F-meas) metric, one can see the ResNet-50, ResNet-18, and MobileNet-v2 presented the best performance compared to other models. Moreover, the VGG-19, FCN-8s, and FCN-16s reported the worst performance compared to other models. For the last metric (MAE), the ResNet-50, MobileNet-v2, and ResNet-18 presented best compared to other models. FCN-8s, VGG-19, and VGG-16 presented the worst compared to other models.

Table 5.4: Comparison of the quantitative scores of different models on ECSSD and MSRA-B.

Note: Higher precision, a larger F-measure, and smaller MAE indicates better performance.

((black bold): good, [red bold]: bad, {blue bold}: good).
















































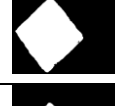

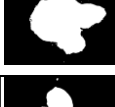













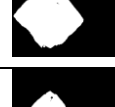













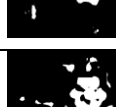













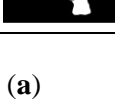




-----	ECSSD				MSRA-B			
Pre-Trained Models	p r e	Recall	F-meas	M A E	p r e	Recall	F-meas	M A E
VGG-16	0.8298	0.4085	0.6501	0.1500	0.8392	0.4962	0.6707	0.1199
VGG-19	0.8334	[0.2943]	[0.5044]	0.1685	0.8664	[0.3647]	[0.4690]	0.1356
ResNet-18	0.8537	0.6782	0.8202	0.0851	0.9389	0.8771	0.9193	0.0307
ResNet-50	0.8705	(0.7969)	(0.8417)	{0.0649}	(0.9394)	(0.9392)	(0.9349)	{0.0183}
MobileNet-v2	0.8218	0.7716	0.7932	0.0800	0.8962	0.8605	0.8811	0.0397
Xception	0.8140	0.6881	0.7687	0.0994	0.8880	0.8316	0.8711	0.0495
InceptionResnet-v2	0.8618	0.6058	0.7466	0.1034	0.9249	0.8427	0.8880	0.0358
FCN-8S	[0.5272]	0.6371	0.5230	0.2176	[0.5823]	0.6396	[0.5665]	0.1795
FCN-16s	0.7904	0.4658	0.6500	0.1450	0.7662	0.5191	0.6585	0.1259
FCN-32s	(0.8913)	0.4062	0.6524	0.1450	0.8690	0.4841	0.7010	0.1167

Table 5.5: Comparison of quantitative scores of different models on DUTS and THUR15k. Note ((black bold): good, [red bold]: bad, {blue bold}: good).





























































































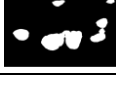



-----	DUTS				THUR15k			
Pre-Trained Models	p r e	Recall	F-meas	M A E	p r e	Recall	F-meas	M A E
VGG-16	0.7616	0.7008	0.7201	0.1140	0.7049	0.5528	0.5811	0.1299
VGG-19	0.7463	0.7049	0.7042	0.1136	0.7625	[0.3850]	0.5074	0.1275
ResNet-18	0.8168	0.7006	0.7657	0.0988	(0.7773)	0.5854	-	0.0984
ResNet-50	(0.8211)	(0.7671)	(0.7912)	{0.0874}	0.7531	(0.8132)	(0.7262)	{0.0901}
MobileNet-v2	0.7516	0.7101	0.7213	0.1125	0.6998	0.7715	0.6678	0.1059
Xception	0.7341	0.6365	0.6787	0.1300	0.7123	0.6650	0.6496	0.1098
InceptionResnet-v2	0.79666	0.5415	0.6751	0.1206	0.7740	0.6982	0.7053	0.0972
FCN-8S	[0.4652]	0.6037	[0.4549]	0.2446	[0.4697]	0.5952	[0.4430]	0.1898
FCN-16s	0.6424	0.5027	0.5677	0.1730	0.6013	0.5476	0.5176	0.1521
FCN-32s	0.7443	[0.4610]	0.6041	0.1568	0.7527	0.4980	0.5903	0.1208

5.5.2 Qualitative Comparison of the State-of-the-Art Pre-Trained Models for SOD

The qualitative results obtained by evaluating the pre-trained models on several example images from ECSSD, MSRA-B, DUTS, and THUR15k datasets are presented in Figure 5.8. As shown, the trained model ResNet-50 performs the best and can detect the most salient object in the image over all datasets (Fig 5.8 a, b).

-----	ECSSD				MSRA-B			
Images								
Ground Truth								
VGG-16								
VGG-19								
(ResNet-18)								
(ResNet-50)								
MobileNet-v2								
Xception								
InceptionResnet-v2								
FCN-8S								
FCN-16s								
FCN-32s								

(a)

-----	DUTS				THUR15k			
Images								
Ground Truth								
VGG-16								
VGG-19								
(ResNet-18)								
(ResNet-50)								
MobileNet-v2								
Xception								
InceptionResnet-v2								
FCN-8S								
FCN-16s								
FCN-32s								

(b)

Figure 5.8: Comparison of salient object detection models on: (a) ECSSD and MSRA-B; (b) DUTS and THUR15k datasets. Note the ResNet-50 produces the best results compared to other models.

5.5.3 Sensitivity Analysis for the Trained Models Against Noise

In this section, we evaluate all trained models for SOD against noise. Three types of noise (salt & pepper, Gaussian, and speckle) were applied to the selected images from two datasets (ECSSD, THUR15K), and the evaluation metrics were calculated. Based on the obtained results, the least sensitive model against noise is ResNet-50 (precision of 0.8739), and the most sensitive model is VGG-19 (precision of 0.7596). Table 5.6 illustrates the evaluation metrics for noisy sample images.

Table 5.6: Evolution metrics of selected trained models for SOD against noise in the selected datasets. Note: (Black bold): good; [Red bold]: bad; {Blue bold}: good).

-----		ECSSD				THUR15k			
Pre-Trained Models	Noise Type	p r e	Recall	F-meas	MAE	p r e	Recall	F-meas	M A E
VGG-16	salt & pepper	0.7939	0.4462	0.5986	0.1451	0.7049	0.5528	0.5811	0.1299
	Gaussian	0.8014	[0.3456]	[0.5453]	0.1645	-	0.3984	-	0.1486
	speckle	0.7818	0.3885	0.5738	0.1586	0.6513	0.4222	[0.4560]	0.1477
VGG-19	salt & pepper	0.8146	0.3695	0.5654	0.1609	0.7625	[0.3850]	0.5074	0.1275
	Gaussian	0.7596	0.3766	-	0.1620	-	0.4370	-	0.1377
	speckle	0.7360	0.3925	0.5663	0.1638	0.6360	0.4693	0.4862	0.1432
ResNet-18	salt & pepper	[0.5324]	(0.8699)	0.5684	0.1997	0.7773	0.5854	-	0.0984
	Gaussian	0.7702	0.6364	0.7131	{0.1136}	0.6374	0.6387	0.5918	0.1156
	speckle	0.7324	0.6461	0.6920	0.1182	[0.5759]	0.6334	-	0.1403
ResNet-50	salt & pepper	0.7571	0.6442	0.6712	0.1324	0.7531	(0.8132)	(0.7262)	{0.0901}
	Gaussian	(0.8739)	0.5576	(0.7292)	0.1146	(0.7585)	0.6256	0.6466	0.1013
	speckle	0.8298	0.5115	0.6761	0.1278	0.6338	0.6277	0.5708	0.1166

5.6 Conclusion

In this work, we evaluated the performance of state-of-the-art deep learning models that employ the semantic segmentation technique for Salient Object Detection. Deep learning models utilize the benefits of neural networks for extracting a hierarchy of features to detect a salient object in an image. We compared the performance of ten state-of-the-art deep learning models over four standard datasets. Based on the evaluation metrics (Precision-

Recall, F-measure, and Mean Absolute Error) of the experimental results from four well-known datasets (ECSSD, MSRA-B, DUTS, and THUR15k), we demonstrate that the Resnet-50 model outperformed the other investigated models, while the FCN-8s model performed the poorest. In addition, we discovered that the ResNet-50 model was the least sensitive against noise, whereas the VGG-16 model was most sensitive. Therefore, the ResNet-50 model offers the most efficient, accurate, and noise-resistant network for Salient Object Detection. As future work, the authors intend to extend this work for the detection of salient objects in video clips. This will require video clips with annotated salient objects and space-time deep learning models.

References

1. Navalpakkam, V. and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). 2006. ieee.
2. Wu, R., Y. Yu, and W. Wang. Scale: Supervised and cascaded laplacian eigenmaps for visual object recognition based on nearest neighbors. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.
3. Ma, Y.-F., et al. A user attention model for video summarization. in Proceedings of the tenth ACM international conference on Multimedia. 2002. ACM.
4. Avidan, S. and A. Shamir. Seam carving for content-aware image resizing. in ACM Transactions on graphics (TOG). 2007. ACM.
5. Zhang, F., B. Du, and L. Zhang, Saliency-guided unsupervised feature learning for scene classification. IEEE Transactions on Geoscience and Remote Sensing, 2014. 53(4): p. 2175-2184.
6. Zhu, J.-Y., et al., Unsupervised object class discovery via saliency-guided multiple class learning. IEEE transactions on pattern analysis and machine intelligence, 2014. 37(4): p. 862-875.
7. Das, A., et al., Human attention in visual question answering: Do humans and deep networks look at the same regions? Computer Vision and Image Understanding, 2017. 163: p. 90-100.

8. Xu, K., et al. Show, attend and tell: Neural image caption generation with visual attention. in International conference on machine learning. 2015.
9. Borji, A., et al., Salient object detection: A survey. Computational Visual Media, 2014: p. 1-34.
10. Long, J., E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
11. Krizhevsky, A., I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. in Advances in neural information processing systems. 2012.
12. Goferman, S., L. Zelnik-Manor, and A. Tal, Context-aware saliency detection. IEEE transactions on pattern analysis and machine intelligence, 2011. 34(10): p. 1915-1926.
13. Harel, J., C. Koch, and P. Perona, Graph-Based Visual Saliency. NIPS. URL <http://papers.klab.caltech.edu/300/1/543.pdf>, 2006.
14. Hou, X. and L. Zhang. Saliency detection: A spectral residual approach. in 2007 IEEE Conference on computer vision and pattern recognition. 2007. IEEE.
15. Jiang, H., et al. Salient object detection: A discriminative regional feature integration approach. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2013.
16. Liu, T., et al., Learning to detect a salient object. IEEE Transactions on Pattern analysis and machine intelligence, 2010. 33(2): p. 353-367.

17. Perazzi, F., et al. Saliency filters: Contrast based filtering for salient region detection. in 2012 IEEE conference on computer vision and pattern recognition. 2012. IEEE.
18. Cheng, M.-M., et al., Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014. 37(3): p. 569-582.
19. Jiang, P., et al. Salient region detection by ufo: Uniqueness, focusness and objectness. in Proceedings of the IEEE international conference on computer vision. 2013.
20. Zhang, J. and S. Sclaroff. Saliency detection: A boolean map approach. in Proceedings of the IEEE international conference on computer vision. 2013.
21. Shen, X. and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. in 2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012. IEEE.
22. Wang, Q., W. Zheng, and R. Piramuthu. Grab: Visual saliency via novel graph model and background priors. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
23. Wei, Y., et al. Geodesic saliency using background priors. in European conference on computer vision. 2012. Springer.
24. Yang, C., et al. Saliency detection via graph-based manifold ranking. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2013.

25. Zhu, W., et al. Saliency optimization from robust background detection. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
26. Li, G. and Y. Yu. Visual saliency based on multiscale deep features. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
27. Wang, L., et al. Deep networks for saliency detection via local estimation and global search. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
28. Zhao, R., et al. Saliency detection by multi-context deep learning. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
29. Lee, G., Y.-W. Tai, and J. Kim. Deep saliency with encoded low level distance map and high level features. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
30. Liu, N. and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
31. Li, G. and Y. Yu. Deep contrast learning for salient object detection. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
32. Kruthiventi, S.S., et al. Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

33. Hu, P., et al. Deep level sets for salient object detection. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
34. Osher, S. and J.A. Sethian, Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. Journal of computational physics, 1988. 79(1): p. 12-49.
35. Mumford, D.B. and J. Shah, Optimal approximations by piecewise smooth functions and associated variational problems. Communications on pure and applied mathematics, 1989.
36. Zhang, P., et al., Salient object detection by lossless feature reflection. arXiv preprint arXiv:1802.06527, 2018.
37. Zhang, L., et al. A bi-directional message passing model for salient object detection. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
38. Garcia-Garcia, A., et al., A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:1704.06857, 2017.
39. Ghariba, B.M., M.S. Shehata, and P. McGuire, A novel fully convolutional network for visual saliency prediction. PeerJ Computer Science, 2020. 6: p. e280.
40. Trembl, M., et al. Speeding up semantic segmentation for autonomous driving. in MLITS, NIPS Workshop. 2016.
41. Yu, Z., X. Wu, and X. Gu. Fully convolutional networks for surface defect inspection in industrial environment. in International Conference on Computer Vision Systems. 2017. Springer.

42. Mohammadimanesh, F., et al., A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem. *ISPRS journal of photogrammetry and remote sensing*, 2019. 151: p. 223-236.
43. Karami, E., M.S. Shehata, and A. Smith, Adaptive polar active contour for segmentation and tracking in ultrasound videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018. 29(4): p. 1209-1222.
44. Simonyan, K. and A. Zisserman, Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
45. Ghariba, B., M.S. Shehata, and P. McGuire, Visual Saliency Prediction Based on Deep Learning. *Information*, 2019. 10(8): p. 257.
46. He, K., et al. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
47. Napoletano, P., F. Piccoli, and R. Schettini, Anomaly detection in nanofibrous materials by CNN-based self-similarity. *Sensors*, 2018. 18(1): p. 209.
48. Ji, Q., et al., Optimized Deep Convolutional Neural Networks for Identification of Macular Diseases from Optical Coherence Tomography Images. *Algorithms*, 2019. 12(3): p. 51.
49. Sandler, M., et al. Mobilenetv2: Inverted residuals and linear bottlenecks. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

50. Chollet, F. Xception: Deep learning with depthwise separable convolutions. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
51. Sifre, L. and S. Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2013.
52. Szegedy, C., et al. Going deeper with convolutions. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
53. Szegedy, C., et al. Inception-v4, inception-resnet and the impact of residual connections on learning. in Thirty-First AAAI Conference on Artificial Intelligence. 2017.
54. Yan, Q., et al. Hierarchical saliency detection. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.
55. Achanta, R., et al. Frequency-tuned salient region detection. in IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009). 2009.
56. Deng, J., et al. Imagenet: A large-scale hierarchical image database. in 2009 IEEE conference on computer vision and pattern recognition. 2009. Ieee.
57. Margolin, R., L. Zelnik-Manor, and A. Tal. How to evaluate foreground maps? in Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

Chapter 6. Conclusions and Future Work

6.1 Conclusions

In this thesis, the capabilities and limitations of deep learning techniques for predicting visual saliency on still images were studied. In Chapter 2, a pre-trained model, (i.e., fine-tuning) was proposed to predict visual saliency. The proposed model used deep learning encoder-decoder architecture based on a transfer learning technique. In particular, the proposed model (VGG-16 model) was trained on a well-known dataset (SALICON) and was tested on several datasets, including TORONTO, MIT300, MIT1003, and DUT-OMRON dataset. The experimental results obtained from the analysis of four datasets explained the superior capability of the proposed model compared to other state-of-the-art models [52].

In Chapter 3, the performance of five state-of-the-art deep learning models (i.e., VGG-16, ResNet-50, Xception, InceptionResNet-v2, and MobileNet-v2) for visual saliency prediction was assessed. These models were trained using the SALICON dataset and then tested over four other standard datasets (i.e., TORONTO, MIT300, MIT1003 and DUT-OMRON). ResNet-50 outperformed the other models and its saliency maps very closely predicted the ground truth data. The poorest performance was observed from the InceptionResNet-v2 model and was likely caused by overfitting due to the large number of layers in this model.

In Chapter 4, a novel Fully Convolutional Network (FCN) was proposed for visual saliency prediction. This model contains three encoder-decoders, three inception modules, and one residual module. The proposed model does not use any pre-trained models; however, this model was trained from scratch and used the data augmentation technique to create variations of images. The experimental results illustrated that the proposed model achieved reasonable results relative to other state-of-the-art models. In this work, increasing the number of training images was found to increase the model prediction accuracy. Furthermore, because the model was trained from scratch, I expected the model will require additional training data, which are currently unavailable as the maximum amount of available training data is currently about 10,000 images (e.g., SALICON).

In Chapter 5, Salient Object Detection (SOD) was employed, because this topic is related to the topic of the thesis (visual saliency prediction). Indeed, SOD consists of identifying a binary map, while the objective of visual saliency is to predict a density map of human eye fixation. In this chapter, the performance of state-of-the-art deep learning models that employ the semantic segmentation technique for Salient Object Detection was evaluated. Deep learning models utilize the benefits of neural networks for extracting a hierarchy of features to detect a salient object in an image. In this work, the performance of ten state-of-the-art deep learning models over four standard datasets were compared. Based on the evaluation metrics (Precision-Recall, F-measure, and Mean Absolute Error) of the experimental results from four well-known datasets (ECSSD, MSRA-B, DUTS, and THUR15k), the Resnet-50 model outperformed the other investigated models, while the

FCN-8s model performed the poorest. Moreover, the ResNet-50 model was found to be the least sensitive against noise, whereas the VGG-16 model was most sensitive.

6.2 Future Work

Although much progress has been made in visual saliency prediction using deep learning techniques, there continues to be room for improvement. For example, all proposed models still fall short of reaching human performance. Therefore, possible recommendations and ways to improve the results and performance of the proposed approaches in future work are highlighted below:

1. The need for higher level visual understanding

Visual saliency models still cannot understand high-level semantic meaning in rich scenes (i.e., the semantic gap). To reach the level of human performance, visual saliency models will need to discover increasingly higher-level concepts in images, including the location of objects, text, and the expected locations of people in images [61].

2. Analysis of evaluation procedures

Most evaluation benchmark metrics are not suitable for all deep models because they are often inconsistent with each other. For example, evaluation metrics for saliency models of video images continue to urge for the design of suitable measures for comparison with deep learning models [62].

3. Collecting high quality data

To improve the performance of visual saliency models based on deep learning, collecting new datasets for training, validating, and testing (still and video images) are required. These datasets are very important and are crucial to the model's progress. Therefore, large scale datasets are useful for training models and achieving high performance. In addition, the analysis of the limitations of models (i.e., weakness) can help contribute to the design of new models, datasets, and applications, for that next qualitative step in performance [61].

4. Salient Object Detection in video images

we suggest extending the work of Salient Object Detection that achieved in chapter 5 for video clips. This work will require video clips with annotated salient objects and space-time deep learning models.

Overall, visual saliency prediction continues to be an important aspect of computer vision and neuroscience fields. Deep learning models, including the novel FCN model proposed in this dissertation, contribute to visual saliency prediction and there is great opportunity for future research in this field.

References

1. Ghariba, B., M.S. Shehata, and P. McGuire, *Visual Saliency Prediction Based on Deep Learning*. Information, 2019. **10**(8): p. 257.
2. Borji, A., *Saliency prediction in the deep learning era: An empirical investigation*. arXiv preprint arXiv:1810.03716, 2018.
3. Borji, A., H.R. Tavakoli, and Z. Bylinskii, *Bottom-up Attention, Models of*. arXiv preprint arXiv:1810.05680, 2018.